Universidad Carlos III Madrid

Máster en Ciencias Actuariales y Financieras

Trabajo Fin de Máster

Modelos de Alerta Temprana: Probabilidades de Impago en el Seguro de Daños (GLM y Machine Learning)

Jaime García Salcedo

04/06/2018

Tutores del Proyecto

José Miguel Rodríguez – Pardo del Castillo

Jesús Ramón Simón del Potro

AGRADECIMIENTOS

Comienzo agradeciendo a mis tutores del proyecto Dr. José Miguel Rodríguez-Pardo del Castillo y Dr. Jesús Simón del Potro por su inestimable dedicación para la correcta consecución del presente estudio, con la aportación de ideas novedosas dada su dilatada experiencia profesional en el mundo actuarial.

Hago extensivo mi reconocimiento a la directora del área de empresas P&C de AXA, Eva Tomás González, por darme la oportunidad de investigar un tema tan apasionante como es el riesgo de impago en el sector asegurador; y a los compañeros del departamento de daños e ingeniería, encabezados por Matías Berestovoy, por la valiosa ayuda brindada para comprender las particularidades que rigen este tipo de contratos.

Una mención especial merecen mis padres Alfonso y Yolanda, que siempre me han apoyado en las decisiones tomadas durante mi trayectoria educativa y en la realización de esta tesis. Hacer extensivo el agradecimiento a mis tíos Gregorio y Cristina y a mi abuela Elena que, junto a mis padres, me inculcaron la cultura del trabajo y el esfuerzo que me ha llevado a la conclusión de este proyecto. Recordar también a los que ya no se encuentran con nosotros, pero que seguro se sentirían orgullosos de ver los conocimientos adquiridos.

Por último, dar gracias también a mi compañero y amigo Alejandro, con el que me sumergí en el estudio de la ciencia actuarial desde el grado en economía; y las grandes amistades encontradas durante el máster en Madeleine y Laura, logrando los tres convertir el estudio en un entretenimiento, enseñándome otras formas de ver la vida y prestando su ayuda en todo aquello en lo que pudiera precisarla.

ABSTRACT

Este documento aborda uno de los principales riesgos a los que se expone una entidad aseguradora: el riesgo de crédito; centrándose en el relacionado con el impago de las primas. Para ello, se ha generado un modelo de predicción, mediante la utilización de una base de datos formada por parte de una cartera de seguros del área de no vida-empresas de una entidad aseguradora, en un período de cinco años. Posteriormente, se introducirán variables macroeconómicas al modelo para comprobar si son significativas y mejoran las predicciones previamente realizadas. Además, se reflexionará sobre posibles medidas que las compañías pueden tomar para reducir el nivel de impagos. Por último, se mostrarán algunas de las técnicas de machine learning que pudiesen ser interesantes abordar en un futuro, en busca de una mejora de la capacidad predictiva de los modelos.

Palabras clave: *Impago de Prima, Riesgo de Crédito, Predicción de Impago, Modelos GLM, Seguro para Empresas*

This paper deal with one of the most important risks an insurance company is exposed: the credit risk; specifically on the non-payment of the premiums. To this end, several types of predictive models has been created, using a database composed by an insurance portfolio in the non-life business area of an insurance company, over a period of five years. Subsequently, some macroeconomics indicators are fitted into the model to evaluate their significance and predictive capacity. In addition, some methods have been considered in order to downgrade premium default rates. Finally, there is an introduction to some machine learning techniques that could be interesting for exploring in the future, following an improvement of models predictive capacities.

Keywords: *Premium Default, Credit Risk, Probability of Default, GLM Models, Machine Learning*

Índice

1. Motivación

Aunque son varios los años ya transcurridos desde una de las mayores crisis económicas de las que se tiene constancia, cuyo origen se sitúa en Estados Unidos con la burbuja inmobiliaria y la crisis de las hipotecas subprime, y que trajo consigo destacables problemas de liquidez y solvencia en numerosas entidades financieras; el análisis de los riesgos a los que se encuentran expuestos las empresas parecen situarse como uno de los principales pilares en el que asentar el plan estratégico hacia el que deben orientar su actividad.

Para ello, las entidades están empleando sistemas cada vez más complejos, que les permitan tener en cuenta el mayor número de escenarios contingentes y los efectos sobre sus activos y pasivos. Esto lo vemos, por ejemplo, en la simulación de estados mediante la aplicación de shocks sobre variables como los tipos de interés o las hipótesis de lapses, contempladas en un contrato de seguro. Las consecuencias de alteraciones de una realidad económica cambiante sobre las finanzas de la compañía pueden ser muy importantes, lo que implicará la necesidad de mostrar una mayor solvencia frente a la incertidumbre futura.

Del mismo modo, existen evaluadores externos conocidos como agencias de rating, que se encargan de analizar los estados financieros de empresas y Administraciones Públicas, otorgándoles una determinada calificación en función de su probabilidad de default, es decir, de que no pueda hacer frente a las deudas contraídas con terceros según las condiciones estipuladas en el contrato. Las más conocidas son Fitch, Standard&Poor´s y Moody´s. Por lo general, tanto Estados Soberanos como empresas con cotización bursátil o que frecuentemente realizan emisiones de títulos de deuda, suelen pagar a estas agencias para que evalúen sus cuentas y les otorguen una nota, puesto que instituciones con una mejor calificación podrán pedir dinero prestado con un tipo de interés más bajo, reduciendo así sus costes de financiación.

Dado que las grandes corporaciones se fijan entre sus objetivos mantener un determinado rating crediticio, resulta fundamental conocer las características de los activos y pasivos que la componen; pues si la entidad ha emitido títulos de deuda con altos tipos de interés, tendrá que pagar más en el momento de su devolución a vencimiento. Por el contrario, si es la empresa la que los ha adquirido, estará asumiendo un mayor riesgo ligado al retorno esperado de la inversión. Un ejemplo muy claro lo

encontramos en el sector bancario con la concesión de préstamos personales, dónde se emplean sofisticados modelos de predicción de impago en función de las características del contratante: profesión, salario, nivel educativo, estilo de vida… Además, se está llevando a cabo un continuo aumento del marco regulador al que los sectores bancario y asegurador están supeditados, como son la normativa de Basilea III y Solvencia II respectivamente, con el objetivo de controlar los niveles de exposición adquiridos por las entidades.

Es en este punto de evaluación crediticia en el que se centra el estudio que se muestra a continuación; pues si bien están muy extendidos en el sector bancario, en el mundo asegurador parece más compleja su modelización. Desde un enfoque amplio, en una compañía aseguradora podríamos diferenciar los departamentos dedicados a "Vida" y "No-Vida" y, dentro del último, cabe distinguir las áreas de "Particulares" y "Empresas".

Introduciéndonos en el departamento de empresas, es práctica habitual en el sector acudir a información de evaluadores externos sobre la calificación de los tomadores, que basan generalmente sus probabilidades de impago en ratios financieros. En cambio, se ha reparado en que puede ser interesante analizar si existen otras vías a través de las que las aseguradoras puedan estimar la calidad crediticia de sus tomadores, atendiendo a las características de la póliza cotizada y no a las finanzas de estos. Este punto puede ser muy relevante para el ramo objeto de análisis, ya que el conseguir una calificación del tomador normalmente implica acudir a un proveedor externo, lo que se traduce en un coste relativamente elevado dado el tamaño de la cartera.

Por ello, se ha considerado realizar un estudio sobre algunas de las variables que definen una póliza en su proceso de cotización y emisión, para comprobar si es posible generar un modelo de predicción de impagos paralelo al tradicionalmente desarrollado con variables financieras por proveedores externos; logrando reducir los costes de la compañía, proporcionando estimaciones de la probabilidad de impago de una nueva póliza y poder tomar, al mismo tiempo, medidas que permitan minorar sus consecuencias, como podría ser la adquisición de un seguro de impago o la realización de recargos sobre la prima técnica en función de dicha probabilidad.

2. Introducción

La Real Academia de la Lengua Española define como seguro al "contrato por el que alguien se obliga mediante el cobro de una prima a indemnizar el daño producido a otra persona, o a satisfacerle un capital, una renta u otras prestaciones convenidas." Dentro de los contratos de seguro los podemos clasificar de la siguiente forma:

a. Vida: también conocidos como "seguros de personas", se caracterizan por tener como objeto asegurado las personas, de manera que las prestaciones dependerán de las contingencias de fallecimiento, invalidez o jubilación de los asegurados, entre otras.
b. No-Vida: se les denomina también "seguros de daños o patrimoniales", pues buscan reparar las pérdidas generadas a consecuencia de un siniestro en el patrimonio del tomador del seguro. En función de este se pueden diferenciar los dirigidos a particulares o a empresas.

Realizado el análisis para una franja temporal de cinco años, para el ramo de "Daños e Ingeniería" del área de "No-Vida, Empresas", se busca obtener un modelo de predicción de impago que permita conocer, en base a la experiencia previa, la probabilidad de que no pague la prima en el futuro una nueva póliza que se está cotizando, al igual que la existencia de posibles factores que puedan alterar este fenómeno durante su estancia en la cartera. Cabe destacar que estos contratos son de tipo Temporal Anual Renovable (TAR), en los que se producirá una actualización automática de la prima, que sólo deberá informarse al cliente en caso de que suponga una subida o bajada superior al 5,00% respecto de la anterior, dos meses antes de su fecha de vencimiento. Además, la renovación de la póliza será inmediata a no ser que el tomador informe de su oposición a la prorroga con un mes de antelación, o de dos meses en caso de ser la entidad aseguradora.

Dado que el ramo objeto de análisis corresponde con el de "Daños e Ingeniería", parece fundamental presentar el tipo de seguros que lo conforman: finalidad de esta modalidad de contratos: metodología seguida para el cálculo de la prima, quiénes conforman las figuras de asegurado y beneficiario, cuáles son las principales coberturas que estos ofrecen,… para, posteriormente, comentar la evolución que los productos de daños han experimentado en cuanto a prestaciones que se adapten a las necesidades del mercado.

El seguro de daños surge con el objetivo de resarcir el daño patrimonial experimentado por el asegurado como consecuencia del acaecimiento de un siniestro. En este sentido, la Ley 50/1980 del Contrato de Seguros en su artículo 26 establece que la tenencia de un seguro de daños, ante la ocurrencia de un siniestro, no puede suponer un enriquecimiento injusto del asegurado; de tal forma que el valor de reposición no deberá ser superior al coste del siniestro, evitando así incrementos en el patrimonio del asegurado tras el evento.

En este caso, la figura del tomador del contrato se corresponde generalmente con la empresa objeto de seguro, cuyo fin no es otro que, ante la ocurrencia de un siniestro, las pérdidas que este le ocasione le sean resarcidas a la mayor brevedad, para poder continuar su actividad y no ver reducido su patrimonio; si bien la forma y cuantía en que será reparado va a estar directamente relacionado con el conjunto de coberturas y condiciones en los que se enmarcaba la póliza contratada

El proceso de cálculo de la prima a cobrar por la entidad aseguradora presenta de una mayor complejidad frente a otros seguros como los de vida u hogar, pues el tamaño de las carteras de los distintos ramos del área empresas son menores, presentan una mayor heterogeneidad en cuanto a los riesgos objeto de seguro, la potencialidad de que ocurra un siniestro y la cuantía que estos pueden suponer (la severidad en este ramo se presenta como una distribución de cola gruesa). Sin embargo, se han establecido una serie de tasas en función de la tipología del riesgo asegurado (capitales, actividad desarrollada y coberturas contratadas), las protecciones que el objeto asegurado posee, junto con otra serie de factores que determinan finalmente la prima técnica. A esta, se le añadirán los descuentos comerciales, gastos, comisiones, impuestos y tributos, para obtener finalmente la prima a cobrar.

Inicialmente, las pólizas de daños estaban diseñadas para dar como cobertura básica la de incendios si bien, podían ser contratadas como complementarias las de caída de rayo y explosión. Estas últimas, con el transcurso de los años, pasaron a formar parte de las coberturas básicas de la póliza, siendo otras las que se empezaron a ofrecer como opcionales, con el fin de hacerlo más atractivo en el mercado. Entre las garantías adicionales, en esta modalidad de seguro, cabe destacar:

- Riesgos extensivos
- Daños eléctricos y por agua

- Pérdidas ocasionadas por robo
- Responsabilidad civil
- Pérdida de beneficios y rotura de maquinaria
- Otras coberturas: asistencia, equipos electrónicos o protección jurídica...

Por tanto, resulta evidente el proceso de modernización que han experimentado los productos de daños para empresas, ofreciendo paulatinamente un mayor número de coberturas, que doten al cliente de las alternativas de contratación necesarias para tener una protección global, aumentando el volumen de negocio generado con su venta.

Inicialmente, se procederá a presentar los datos empleados en el estudio y el tratamiento que se ha realizado sobre ellos, junto con las principales variables propias de la póliza cotizada que serán objeto de análisis, para continuar con un análisis descriptivo de las mismas y la relación que a priori deberían tener con la probabilidad de impago.

A continuación, se llevará a cabo la estimación de distintos modelos de regresión, que permitan descartar aquellas variables que no tengan un impacto significativo sobre el impago. Además, se establecerá un doble análisis de predicción en una muestra de validación, que permita visualizar la calidad de los modelos, para distintos niveles de probabilidad a partir de los que considerar este fenómeno, junto con la creación de un rating de crédito como herramienta para clasificar las pólizas, en función de las predicciones realizadas, en línea con los modelos bancarios.

Posteriormente, se incluirán también en el estudio algunos indicadores de índole macroeconómica, para comprobar si las predicciones mejoran con su inclusión, tanto en términos de significatividad como en reducción de impagos no previstos. El proceso de validación para estas será similar al comentado en líneas anteriores, posibilitando así la comparación y análisis de las mejoras que producen su inclusión.

La finalidad, por tanto, es reducir la exposición de la compañía al riesgo de crédito que supone el impago de la principal fuente de ingresos de las entidades aseguradoras: las primas. De igual forma, servirá para dotar de una mayor o menor flexibilidad al cliente a la hora de contratar la póliza, puesto que, si se estima que una nueva cotización tiene una alta probabilidad de impago, se podrían establecer reglas de validación que no le permitan seleccionar la modalidad de pago fraccionado o imputar un spread de crédito sobre la prima técnica que le correspondería pagar o cancelar su emisión.

3. Desarrollo

Fuente de Información

Para la realización de un estudio del que poder extraer conclusiones relevantes y sobre todo, consistentes con la realidad del mercado asegurador, resulta fundamental la tenencia de una base de datos que se comporte cómo la rama objeto de análisis; siendo en este caso la correspondiente al seguro de daños para empresas. Con el objetivo de que sea una imagen fiel de una cartera real, pero no implique la revelación de información confidencial sobre la entidad de la que se han empleado los datos, se han introducido factores y tasas en algunas variables.

Dado el gran número de variables que se quieren contemplar en el análisis, resulta fundamental la formación de una base de datos lo suficientemente grande y completa, lo que implica la necesidad de utilización de distintas fuentes y repositorios de información. Para ello, ha sido necesaria la búsqueda, selección, filtrado y combinación de aquellas cuestiones de interés, que generalmente suelen estar sin depurar. Si bien la mayor parte de los datos (en bruto) que se han utilizado son internos, a la hora de contemplar cuestiones de índole macroeconómica se ha acudido al Instituto Nacional de Estadística, pues se presenta como la principal fuente de información para cuestiones relacionadas con la economía nacional (PIB, tasas de paro,…).

Las herramientas que se han empleado para las labores mencionadas son las siguientes:

- SQL: software cuyo potencial reside en la facilidad para realizar consultas en grandes bases de datos, permitiendo el cruce entre distintos repositorios de información. Aunque son numerosas sus funcionalidades, se ha utilizado fundamentalmente para dotar de formatos similares a la totalidad de variables, extraer la información de las pólizas y organizarlas, evitando duplicidades, teniendo siempre en consideración para su selección el último año en que estuvieron en vigor, para tener los datos lo más actualizados posibles.
- Excel: diseñada como hoja de cálculo, es uno de los softwares más extendidos en los usos universitario y profesional, al permitir la realización de numerosas funciones matemáticas, generación de gráficos e incluso, de calculadoras para tarificación y simulación a través de su herramienta de programación “Visual Basic for Applications”, muy extendida en la práctica actuarial. Se ha utilizado principalmente para la realización de recuentos y gráficos, cálculos sobre

estadística descriptiva y apoyo para la comprobación del correcto estado de los datos, antes de ser leídos para la modelización con el paquete estadístico R.

Presentación de Variables

Definiendo el impago de prima como la situación por la que el tomador del seguro no hace frente, en los términos de cuantía y plazo establecidos en la póliza, a los pagos de la prima del seguro y que supone, tras un mes como plazo para ser satisfecha, la inmediata extinción de la obligación contractual de la entidad aseguradora de cubrir los daños que pudiera sufrir el riesgo asegurado, al implicar su anulación; se presenta la variable objeto de análisis como un factor binario que puede tomar los siguientes valores:

$$\boldsymbol{Impago\_Prima(Y)} = \begin{cases} \mathbf{0} & \boldsymbol{la\ Póliza\ Paga\ la\ Prima} \\ \mathbf{1} & \boldsymbol{la\ Póliza\ NO\ Paga\ la\ Prima} \end{cases}$$

Dado que el impago de la prima implica la anulación de la póliza, pero esta relación no se da de igual forma en sentido inverso, al existir numerosos motivos de anulación de los contratos de seguro más allá de la falta de pago como son la decisión unilateral o fallecimiento del tomador, la desaparición del riesgo u otras; se ha generado una variable que capture la situación de la posible vigencia de la póliza o el motivo de su anulación, organizada esta última en dos grupos, como se expresa a continuación. Este factor puede resultarnos de gran utilidad para comprobar el peso que tiene el impago como motivo de anulación de las pólizas que componen la cartera empleada para el análisis.

$$\text{Vigor\_Anulada(X80)} = \begin{cases} 0 & \text{Póliza en Vigor} \\ 1 & \text{Anulación por Impago} \\ 2 & \text{Anulación por Otro Motivo} \end{cases}$$

Una vez presentada la variable de interés sobre la que se sustentarán los modelos que se desarrollarán en las próximas hojas, resulta imprescindible definir aquellos factores que introduciremos en el análisis y las características a nivel cualitativo y cuantitativo que albergan; pues la estimación de un modelo precisa tanto de una variable respuesta a predecir, como de variables de las que nutrirse, denominadas explicativas.

Como mencionaba en líneas anteriores, el proceso de modelización quedará estructurado en dos fases: la primera, que contendrá únicamente aquellas características propias de cada póliza y que determinarían (en términos brutos) la prima a cobrar según su riesgo, coberturas y procedencia; para posteriormente, incluir indicadores

macroeconómicos que intuitivamente mejorarán la calidad de las predicciones iniciales. Además, serán analizadas desde una óptica teórica algunas variables que pudiesen tener un efecto significativo sobre el impago si bien, dado que se carece de datos para gran parte de la cartera, no se tendrán en cuenta para la modelización.

Variables propias de la póliza

Quedan amparadas bajo este concepto todo factor que podríamos considerar como característico de una póliza y que, de forma automática, no quedaría replicado en otra en el momento de su emisión. A continuación quedan definidas aquellas que introduciremos en los modelos:

- X4: Familia Sectorial que establece, en función de la actividad empresarial a la que se dedique la entidad, el sector económico (primario, secundario o terciario) al que se encuentra asociado. Aunque la cantidad de actividades económicas que componen la cartera es muy amplia y diversa, ha sido posible su clusterización en dos etapas mediante el empleo de variables intermedias que las aglutinen:
    - En función de su "familia", que representa la actividad a la que se dedica la empresa, desde una óptica amplia. A modo de ejemplo, si son actividades relacionadas con invernaderos o explotaciones agrícolas de cereales, estas tendrían un identificador común (familia) definido como "Agricultura".
    - En función de su sector económico, en el que las actividades del ejemplo anterior que pertenecían a la familia de agricultura, junto con otras del mismo sector productivo, como las integrantes de la familia de "Ganadería", quedarán recogidas en el sector primario (S1).

- X6: Mes de origen en que la póliza entró en vigor, y a partir del cual se han generado otras dos variables, para analizar la posible existencia de dinámicas de consumo y contratación:
    - X94: recoge de forma binaria si la póliza se originó en el mes de diciembre o en cualquier otro mes del año. El objetivo es analizar si las empresas que contratan pólizas en el mes de diciembre tienen una mayor probabilidad de no pagar la prima.

- X9: Capitales asegurados por la póliza, sobre los que se han aplicado logaritmos para reducir su dimensión. Para la cartera de seguros objeto de análisis, lo conforman principalmente:
    - Valor del continente: compuesto por edificios, locales, obras de reforma, muros y vallas; esto incluye las construcciones principales y accesorias en las que la empresa lleva a cabo la actividad declarada en la póliza.
    - Valor del contenido: formado esencialmente por el mobiliario profesional y las existencias necesarias para que la entidad desarrolle su actividad. Además se incluyen los vehículos en campa exterior, es decir, aquellos vehículos que se encuentran fuera de un recinto cubierto pero dentro del espacio asegurado.

- X11: variable binaria que informa de la existencia o no de coaseguro en la póliza; entendiendo por coaseguro la situación por la cual varias empresas se reparten un porcentaje del riesgo total de la póliza de forma que, ante un siniestro, cada una de las entidades aseguradoras dentro del coaseguro tendrá que asumir (de forma independiente) el porcentaje del coste del siniestro acordado. A priori, cabría pensar que pólizas con coaseguro deberían tener una probabilidad de impago menor, ya que el número de empresas involucradas es más elevado.

- X12: recogen, en forma de variable binaria, si las pólizas tienen o no reaseguro; es decir, si todo o parte del riesgo de la póliza es transferido a otra u otras entidades de seguro a cambio del pago de una prima.

- X14: representa la prima neta asociada a cada póliza, que se calcula en base a los capitales asegurados, coberturas contratadas y las características del riesgo (actividad, localización…) incluyendo los recargos por gastos, comisiones, impuestos y tributos; y sobre la que se han aplicado logaritmos para reducir su dimensión. Cabe pensar que la probabilidad de impago será más alta cuanto mayor sea la prima que tenga que satisfacer el cliente.

- X16: Variable binaria que recoge si la póliza, en el último año en que estuvo en vigor, era o no de nueva producción, es decir, si acababa de entrar en cartera o llevaba, en el momento de anulación, más de un año en ella.

- X20: factor que clasifica las pólizas en función de su antigüedad y de la que cabe esperar que la probabilidad de impago se reduzca cuantos más años lleve en la cartera. Es interesante su inclusión para dotar de dinamismo al modelo, ya que permitirá evaluar la póliza en los distintos años en que se encuentre en la cartera. Se han agrupado de la siguiente forma:

$$\text{Antigüedad(X20)} = \begin{cases} 1 & \text{1 Año o Menos en Cartera} \\ 2 & \text{Entre 1 y 5 años en Cartera} \\ 3 & \text{Más de 5 años en Cartera} \end{cases}$$

- X22: Recoge como factor el canal de distribución a través del cual ha generado la póliza, realizando una distinción especial entre corredores (C) y brókers (BR) de seguros, que se definen como empresas de intermediación que venden seguros de varias compañías a cambio de una comisión; agentes (A), que comercializan únicamente seguros de la empresa a la que pertenecen, negocio directo (D) comercializado por internet y red telefónica por la propia empresa; y otros canales (O) distintos a los anteriores.

- X23: Clasifica las pólizas, a través de un código numérico asignado en función de la zona geográfica en la que se encuentran el tomador del seguro, y de la que se espera presenten distintos niveles de impago.
  Estas regiones son: Norte (1), Este (2), Levante (3), Centro (4), Sur (5), Oeste (10) y Servicios Centrales (9)

- X27: Corresponde con la periodicidad con la que se satisface la prima, que puede ser de carácter Anual (A), Semestral (S), Trimestral (T) o Irregular (I). A través de esta variable se ha generado una nueva (X28) de carácter binario que toma los valores Único (U) si la forma de pago era Anual, y Fraccionado (F) en caso de que fuese Semestral, Trimestral o Irregular. Desde una óptica teórica, cabe pensar que la probabilidad de impago guarda una estrecha relación con su periodicidad, puesto que, al ser pólizas del tipo temporal anual renovable, si la prima está fraccionada, existen más incentivos del tomador a no hacer frente a

alguno de sus pagos mientras que la cobertura, hasta el momento en que impague, sería total.

- X29: Informa si la prima se satisface mediante domiciliación bancaria o es el tomador el que acude a su agente o mediador a realizar su abono. A priori, la probabilidad de impago debería ser menor si la póliza no está domiciliada, al implicar una relación más cercana entre tomador y mediador.

$$\text{Domiciliación(X29)} = \begin{cases} 0 & \text{la Póliza no está Domiciliada} \\ 1 & \text{la Póliza está Domiciliada} \end{cases}$$

- X31: Factor que indica el tramo por el que fue emitido la póliza, realizando una distinción entre el automatizado, si el riesgo tenía agravación normal y suma asegurada igual o inferior a tres millones, y especializado, en caso de que fuera un riesgo agravado o excluido o la suma asegurada fuera superior a tres millones aun siendo un riesgo normal.

$$\text{Tramo(X31)} = \begin{cases} 0 & \text{Automatizado} \\ 1 & \text{Especializado} \end{cases}$$

  La diferencia existente entre tramos es la delegación de las labores de suscripción que se realiza en el tramo automatizado hacia los mediadores, frente al tramo especializado en el que un suscriptor experimentado debe liberar la póliza para permitir su emisión, tras comprobar el objeto del riesgo y las garantías que recoge.

- X33: Variable binaria relacionada con el área de distribución, que informa si la persona que ha suscrito la póliza pertenece o no a un determinado grupo de mediadores o agentes cuyos conocimientos técnicos se entienden como mejores y que, además, presentan unos niveles de producción y siniestralidad fijados desde el ramo. Teóricamente, los niveles de impago deberían ser menores para pólizas cuyos mediadores se encuentran en este grupo.

$$\text{Club\_Mediadores(X33)} = \begin{cases} 0 & \text{Mediador No Pertenece al Club} \\ 1 & \text{Mediador Pertenece al Club} \end{cases}$$

- X72: Recoge el número de siniestros que ha sufrido la entidad asegurada durante los últimos años. Este dato puede tener dos fuentes de origen para su correcta modelización:

- Póliza de Nueva Producción: esta información debería declararla el tomador del seguro a la hora de realizar la cotización y contratación de la póliza de forma veraz, ya que, en caso contrario y si aconteciera un siniestro, la póliza podría declararse nula por estar el contrato "viciado" (deber del asegurado de declarar correctamente el riesgo asegurado).
- Póliza en Cartera: recogida por el departamento de siniestros de la propia entidad en base a la información histórica de la que se dispone.

Variables macroeconómicas

Esenciales a la hora de entender el comportamiento empresarial, se presentan como indicadores de la actividad económica nacional, tanto a nivel estructural como coyuntural. A tal efecto, se han considerado las siguientes variables para completar el modelo:

- X89 y X90: corresponden con la Tasa de Paro y la Variación del Paro registrado, quizás las variables que mejor expliquen la situación económica que atraviesa la región, pues cuanto mayor sea la tasa de paro mayor debería ser la probabilidad de impago, mientras que valores negativos de su variación expresarían una mejora de la situación económica, lo que reduciría el número de impagados.

- X91, X92 y X93: contienen la información referente a la estructura sectorial regional, en términos porcentuales, respecto a su PIB total, es decir, el peso de los sectores Primario (Agricultura, Ganadería, Pesca y Explotación Forestal), Secundario (Industria, Energía y Extracción de Materias Primas) y Terciario (Servicios Empresariales). En este caso, no es del todo claro su efecto sobre el impago ya que, si bien el sector servicios es un buen indicador del desarrollo económico, y en las zonas más desarrolladas estas tasas de impago deberían ser menores, las actividades que componen este sector están muy expuestas a la coyuntura económica, teniendo un efecto contrario al mencionado. En caso del peso del sector secundario, quizás deberían mostrar menores probabilidades de impago al entenderse como regiones con alto nivel de industrialización y un empleo, generalmente, más estable.

Dado que el período de análisis es superior al año, se ha considerado analizar el valor que tomaban estos indicadores en el último año en que cada póliza estuvo en cartera, es decir, si la póliza se originó en 2014 pero fue en el 2016 en el que se produjo su impago/anulación, el valor considerado para la variable será el del año 2016. Del mismo modo, y dado que se dispone de la información provincial del tomador, estas variables se encuentran organizadas tanto en función de su último año en vigor, como de su situación provincial. De esta forma conseguiremos extraer con más precisión el impacto de la coyuntura económica (a través del paro o su tasa de variación) y la estructura sectorial en las probabilidades de impago de la prima.

Otras variables que considerar

Dado el elevado volumen de pólizas que componen una cartera de seguros, y el frecuente aumento de variables que en estas se suelen considerar, encontramos que numerosas pólizas presentan carencias de datos para factores que, a priori, pueden guardar un efecto significativo en la explicación del impago de la prima. Por ello, con el objetivo de evitar sesgos de muestreo, las siguientes variables no serán incluidas en la modelización, al estar informadas únicamente para el último año:

- X87: Corresponde con la dimensión de la entidad asegurada, basándonos en el régimen estatutario en el que se constituyó y el número de empleados que posee. A tal efecto, se ha considerado la siguiente clasificación:

$$\text{Dim\_Empresa}(X87) = \begin{cases} 0 & \textit{Sin Informar} \\ 1 & \textit{Autonónomos y Pymes} \\ 2 & \textit{Grandes Empresas y Corporaciones} \\ 3 & \textit{Administración Pública y Comunidades de Propietarios} \end{cases}$$

  Intuitivamente, cabe pensar que empresas de una mayor dimensión tengan una probabilidad de impago más pequeña frente a autónomos y pymes, que se encuentran más expuestas ante los ciclos económicos y el impacto sobre sus resultados.

- X84: Registra de forma binaria si la póliza, antes de su emisión, tuvo que ser validada por un suscriptor más experimentado, bien sea por la complejidad del riesgo asegurado o por las coberturas que esta cubre:

$$\text{Proceso\_Validación}(X84) = \begin{cases} 0 & \text{No Pasó por Proceso de Autorización} \\ 1 & \text{Pasó por Proceso de Autorización} \end{cases}$$

- X32: Dota de una valoración al cliente en función del número de pólizas que tenga contratadas con la entidad, su siniestralidad esperada y las primas que se tienen previsto cobrar durante su vigencia. A tal efecto, definiríamos a los clientes tipo "A" como los más rentables, seguidos de los clientes "B" y "C" indistintamente y siendo los tipo "F" y "G" los menos rentables para la entidad. De esta forma, clientes con una mejor calificación deberían mostrar menores probabilidades de impago, aunque podrían darse comportamientos asimétricos por el coste del seguro o una mala experiencia con otras pólizas contratadas con la compañía.

- X34: Clasifica las pólizas en función del riesgo que supone para la entidad aseguradora la actividad declarada por el asegurado; realizando una distinción entre:

$$\text{Agravación}(X34) = \begin{cases} \text{N} & \textit{Riesgo Normal} \\ \text{A} & \textit{Riesgo Agravado} \\ \text{AP} & \textit{Riesgo Agravado} + \\ \text{E} & \textit{Riesgo Excluído} \end{cases}$$

Análisis Descriptivo

Una vez presentadas las variables que conforman la base de datos, se ha llevado a cabo un análisis de carácter univariante entre el impago de la prima y el resto de los factores que se consideran en los modelos que se recogen en líneas posteriores. El objetivo es poder avistar cómo se distribuye el impago en función de cada una de ellas, teniendo así una primera aproximación a la variable de interés.

Para ello, y dado que estamos tratando con una variable objetivo de carácter binario (Pago/Impago) que podríamos entender como una probabilidad cuyos valores siempre han de estar comprendidos entre cero y uno, emplearemos a modo de apoyo un sistema de recuento del número de observaciones que, dada una situación de Pago/Impago poseen una determinada característica o pertenecen a un grupo específico, como se muestra en el siguiente esquema:

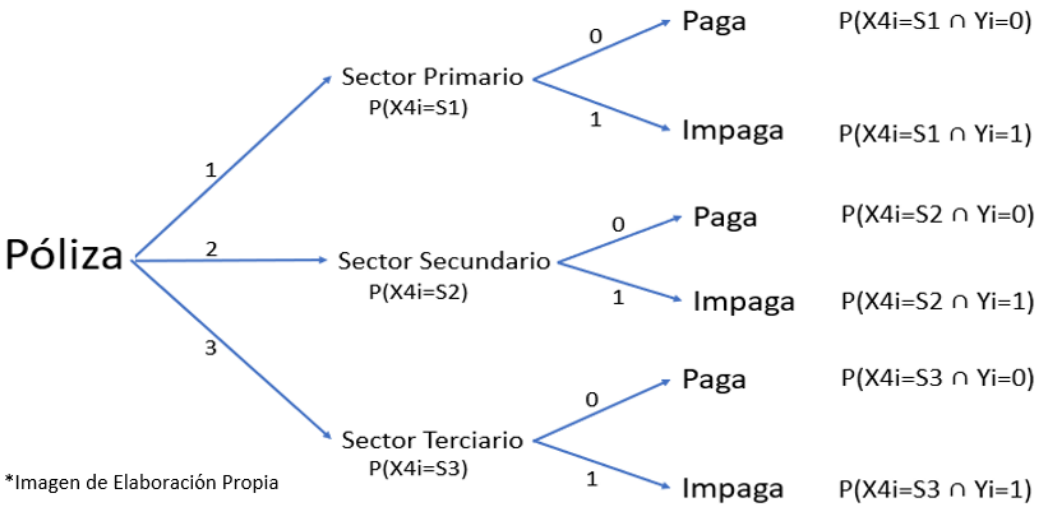


*Imagen de Elaboración Propia

La imagen anterior es útil para tomar una visión inicial de posibles comportamientos hacia el impago, en función de la pertenencia o no a una determinada clase, en aquellas variables que se presentan como categóricas o, en caso de que fueran continuas, podrían agruparse en base a algún criterio, por ejemplo, estableciendo tramos de prima o capital; posibilitando también su realización.

Para este acercamiento previo a la estimación de modelos, nos apoyaremos en una de las principales ramas de las matemáticas: la teoría de la probabilidad. Esta se encarga del estudio de los fenómenos aleatorios, es decir, analizan como los resultados que puede obtener una variable de interés pueden ser diversos, pese a tener las mismas características iniciales; en el ejemplo de la imagen notamos la existencia de pólizas que pagan la prima junto a otras que no, tanto si su actividad empresarial está orientada al Sector Primario (X4=S1), como en caso de pertenecer a los Sectores Secundario (X4=S2) o Terciario (X4=S3); sin embargo, en un caso determinista, si una póliza tuviese una determinada característica, podríamos inducir de forma directa que va, por ejemplo, a no pagar la prima con una probabilidad del 100%, pues no existiría un componente aleatorio por el que pudiese tomar otro valor (que pagase), aunque su probabilidad fuese ínfima. En este punto, es momento de enunciar el Teorema de la Probabilidad Total y el Teorema de Bayes, empleados para los cálculos de los resultados univariantes que se muestran a continuación.

- Teorema de la Probabilidad Total:

  "Sea A1, A2, …, A_n una partición sobre el espacio muestral y sea B un suceso cualquiera del que se conocen las probabilidades condicionales $P(B|A_i)$, entonces la probabilidad del suceso B viene dada por la expresión"[1]:

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

  Esta expresión relaciona la probabilidad de ocurrencia de un suceso como el sumatorio de las probabilidades de ocurrencia de este, en cada uno de los caminos por los que puede producirse. En el ejemplo del esquema anterior, la probabilidad total de impago, $P(Y = 1)$, sería la siguiente:

$$\boldsymbol{P(Y = 1)} = \sum_{i=1}^{3} P(I\,|X4_i) \cdot P(X4_i)$$

$$\boldsymbol{P(Y = 1)} = P(Y = 1 \cap X4 = 1) \cdot P(X4 = 1) + P(Y = 1 \cap X4 = 2) \cdot$$
$$\cdot P(X4 = 2) + P(Y = 1 \cap X4 = 3) \cdot P(X4 = 3)$$

  Es decir, correspondería con la suma de las probabilidades de impago si la actividad de la póliza se enmarcaba en el sector primario, o en el sector secundario o en el terciario.

- Teorema de Bayes:

  Enunciado por Thomas Bayes en 1763, relaciona la probabilidad condicional de un evento aleatorio A dado B en función de la distribución de probabilidad de B dado A y sus funciones de distribución marginales de A.

  "Sea A1, A2, …, A_n un conjunto de sucesos mutuamente excluyentes y exhaustivos, y tales que la probabilidad de cada uno de ellos es distinta de cero. Sea B un suceso cualquiera del que se conocen las probabilidades condicionales $P(B|A_i)$; la probabilidad $P(A_i|B)$ viene dada por"[2]:

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{P(B)}$$

[1] https://es.wikipedia.org/wiki/Teorema_de_la_probabilidad_total
[2] https://es.wikipedia.org/wiki/Teorema_de_Bayes

Siendo $P(B)$ la probabilidad total de ocurrencia del suceso B vista en líneas anteriores y $P(B|A_i)$ la probabilidad de que suceda el evento B dado que ha sucedido A_i.

Continuando con el ejemplo anterior, en el que habíamos obtenido la expresión para la probabilidad total de impago $P(Y = 1)$, si quisiéramos conocer la probabilidad de que sabiendo que no ha pagado la prima, su actividad empresarial estuviera enmarcada en el sector primario, la forma de calcularlo sería la siguiente:

$$P(X4 = S1|Y = 1) = \frac{P(Y = 1|X4 = S1) \cdot P(X4 = S1)}{P(Y = 1)}$$

A través de estas herramientas estadísticas, pero usando un método de organización inverso, es decir, diferenciando en primer lugar las pólizas que pagan la prima $(Y = 0)$ y las que incurrieron en impago $(Y = 1)$ y en la segunda etapa, cada una de las variables objeto de análisis para ambas situaciones, obteniendo los siguientes resultados[3]:

- Las actividades relacionadas con el sector terciario tienen aparentemente unas tasas de impago superiores a las de los sectores primario y secundario, posiblemente asociado a que los negocios que se ubican en el sector servicios suelen tener verse avocados a cesar su actividad si no consiguieron penetrar con éxito un mercado dominado por grandes corporaciones y otras empresas ya establecidas. Además, presentan un mayor grado de adaptabilidad para el traspaso o cambio de actividad de negocio, frente a las actividades de los otros sectores, que suelen precisar de grandes inversiones iniciales y que ofrecen una menor capacidad de metamorfosis.

  Esto es importante ya que el peso de las pólizas sobre actividades del sector terciario es muy destacable, dado que tratamos con seguros sobre empresas, y la estructura económica del país es fundamentalmente terciaria.

[3] Por motivos de confidencialidad, se mostrará únicamente la intuición de los resultados obtenidos

- Analizando el comportamiento mensual del impago se observa, a la luz del siguiente gráfico, una diferencia de más de dos puntos porcentuales entre las tasas máximas y mínimas, al igual que parecen existir unas mayores tasas de impago entre los meses de abril, mayo y junio, a las que le sigue un descenso que se acentúa en los últimos meses del año.

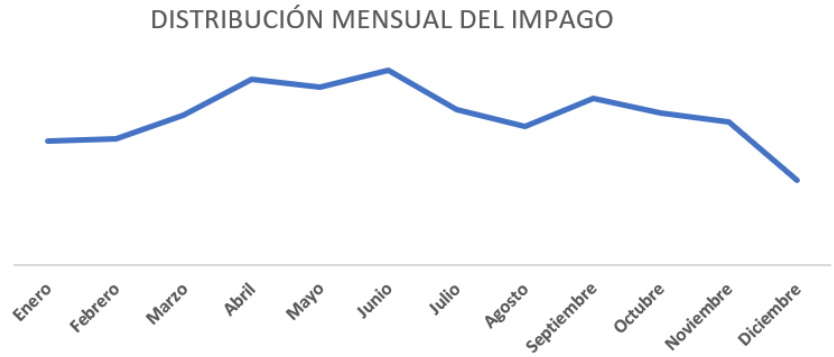


*Gráfico de Elaboración Propia

Aunque inicialmente se habían considerado otros criterios de caracterización de observaciones en base al mes en que se originaron las pólizas, parece recomendable emplear otro criterio: el trimestre en que estas fueron contratadas (X96), pues parece existir una estructura tendencial en casi todos los trimestres, con crecimiento en el primero, un repunte en el segundo que se compensa con una leve disminución en el tercero, siendo el último trimestre en el que se minimizan las tasas de impago.

- Atendiendo a la existencia de coaseguro, observamos una clara distinción entre el comportamiento de las pólizas que se encuentran bajo este régimen y las que no, dónde los niveles de impagados para pólizas con coaseguro son inferiores al cinco por ciento, lo que es bastante significativo, sabiendo además que estas aumentan de forma considerable en caso de que no tengan coaseguro. Cabe destacar que esta modalidad de contratación suele darse para asegurar, principalmente, grandes riesgos o actividades muy agravadas, para reducir el impacto de un posible siniestro grave al distribuir el coste en la proporción acordada entre las entidades coaseguradoras, lo cual suele ir asociado a unos criterios de concesión del seguro mucho más estrictos.
- En cuanto a la posible tenencia de reaseguro, se produce un efecto bastante similar al coaseguro, es decir, existen diferencias notables entre las pólizas con y sin reaseguro, si bien los casos de impago son levemente superiores en este caso.

- Analizando la diferencia entre las pólizas que conformaban la cartera frente a las de nueva producción, encontramos como las últimas poseen una probabilidad de dos puntos porcentuales superior a las primeras en términos de impago, lo cual denota un claro efecto temporal.
- Siguiendo con lo anterior, podemos profundizar más en este efecto si atendemos a la antigüedad de las pólizas, dónde quedarían enmarcadas las de nueva producción en aquellas con valor en $X20 = 1$. Con este nuevo análisis, encontramos que la tasa de impago para pólizas con antigüedad igual o menor a un año es bastante superior a la tasa promedio, si bien esta disminuye de forma muy notable si la póliza tiene una antigüedad de entre dos y cinco años y en más del cincuenta por ciento respecto a las primeras, para las que llevan más de cinco años en cartera.
- En cuanto al canal de emisión encontramos como son las pólizas emitidas por miembros de la propia entidad las que poseen mayores niveles de impago, frente a las que proceden de corredores y brókers. De igual modo, son los últimos en los que se encuentran tasas menores, probablemente asociado a que manejan grandes cuentas por las que obtienen una comisión según la prima y, en caso de impago, esta la perderían.
- Como podemos observar en el siguiente gráfico, son las zonas geográficas del sur y centro de España las que presentan unos mayores niveles de impago, concentrando cerca del 40% del total, frente a regiones del norte, este y oeste de nuestra geografía, con tasas notablemente inferiores. Los servicios centrales son un caso especial, ya que su volumen de producción es bajo y suelen emitir pólizas muy particulares.

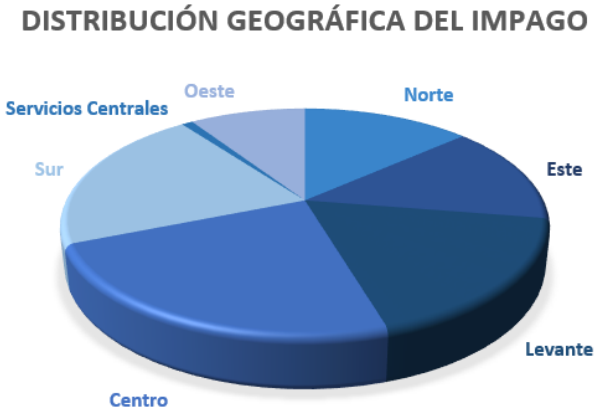


*Gráfico de Elaboración Propia

- En referencia a la periodicidad de pago, hay que recordar que se poseía una variable que las clasificaba en función de si este era de carácter anual, semestral, trimestral o irregular y en la que observamos como son las pólizas con pago semestral las que mayores tasas de pago presentan seguidas de las trimestrales, encontrando una diferencia entre las tasas anuales y semestrales próximas a los diez puntos porcentuales.

  Por ello, y dadas las grandes diferencias encontradas entre las distintas modalidades de pago, se generó la variable X28 que agrupa las observaciones en función de si este se realizaba de forma anual o fraccionada; obteniendo que, ciertamente, el impago es notablemente superior en el caso de las fraccionadas.
- Atendiendo a la domiciliación del pago, y en línea con lo previamente mencionado, son las pólizas no domiciliadas las que presentan menores tasas de impago, siendo además importante la diferencia frente a las que si se encuentran asociadas a una cuenta bancaria. Este hecho puede ser causado por la relación de confianza que generalmente existe entre tomador y agente.
- Con relación al tramo en que se produce la póliza, no encontramos diferencias sustanciales entre la entrada por el automatizado o por el especializado, si bien en el primero los impagados son ligeramente superiores. Destacar también que, dada la existencia de pólizas con el campo sin informar, se han calculado en base a las que sí lo poseían.
- La variable que también parece influir en las probabilidades de impago es el tipo de mediador de la póliza, en la forma en que este se encuentra dentro del club de empresas que, recordando lo comentado, hacía referencia a un grupo exclusivo de mediadores que por volumen de producción y resultados se les había dotado de una denominación y condiciones especiales. A tal efecto se observa una diferencia de cuatro puntos porcentuales en favor de los pertenecientes a este colectivo.

Aunque se va a trabajar con más variables de las que hemos realizado este análisis previo, como son los capitales asegurados y las primas, puede ser relevante mencionar cómo las pólizas que han sufrido algún siniestro presentan unas tasas de pago de prima superiores a las que no lo han tenido.

Modelización

Una vez realizado esta breve introducción de la distribución del impago, en función las características de las pólizas, se va a examinar como actúan de forma simultánea; pues el análisis inicial, al ser univariante, muestra información sesgada de la significatividad que los factores introducidos pueden tener en la explicación de impago.

Para la realización de este nuevo análisis, de carácter multivariante, haremos uso del software libre "R Studio", con el que se podrán estimar distintos modelos de regresión para comprobar el impacto y significatividad que las variables presentan a la hora de explicar el impago de la prima. Posteriormente, se hará uso de los modelos para realizar predicciones sobre esta probabilidad, con el objetivo final de verificar si los valores estimados se comportan como los observados.

La Distribución del Impago

Antes de adentrarnos en los distintos métodos de modelización disponibles, su forma de implementarlos y el motivo por el cual se ha considerado utilizar los recogidos en líneas posteriores, frente a otros de menor complejidad de cálculo e interpretación, parece fundamental analizar en profundidad la variable que se busca explicar con estos modelos: el impago de la prima.

Como se comentaba al inicio, esta variable recoge si las pólizas pagaron o no la prima según lo estipulado en el contrato, por lo que es de tipo binario, es decir, con la información que se posee podemos determinar si la póliza incurrió en impago o si, en caso contrario, cumplió con los pagos en los momentos acordados, tomando los valores $Y = 1\ e\ Y = 0$ respectivamente. De tal forma, ex – ante se puede decir que una póliza tiene durante su vigencia una probabilidad de pago/impago asociada.

En línea de lo recogido en el campo de la estadística, por la teoría de la probabilidad, este tipo de evento dicotómico se comporta como una distribución de Bernoulli $X \sim Be(p)$, dónde "X" representa a la variable aleatoria del experimento y "p" el parámetro con que se distribuye esta variable aleatoria, correspondiendo con la probabilidad de éxito del evento, es decir, que la variable X tome el valor 1; siendo "$1 - p$" la probabilidad de lo que se considera fracaso en el experimento. A continuación, se muestran su función de probabilidad, formulación y sus principales propiedades:

- **Función de Probabilidad:** $f(x) = p^x \cdot (1-p)^{1-x}$; dónde X=0,1
- **Fórmula:** $f(x;p) = \begin{cases} p & si \quad x = 1 \\ q & si \quad x = 0 \end{cases}$
- **Esperanza Matemática:** $E[X] = p$
- **Varianza:** $Var[X] = p \cdot (1-p) = p \cdot q$
- **Función Generadora de Momentos:** $mgf = (q + pe^t)$

Atendiendo a esta formulación, la probabilidad de impago de una póliza correspondería con que la variable aleatoria tomase el valor 1 (éxito), es decir, correspondería con "p". Dado que los casos en que $Y = 1$ son una minoría, respecto a la totalidad de pólizas que componen la cartera ($p < q$) podemos concluir que la moda de la variable aleatoria será igual a cero.

A parte de las propiedades de la distribución de Bernoulli que previamente se han mencionado, esta cobra mayor importancia dentro del estudio si atendemos a su aproximación a la distribución binomial, también de carácter discreto y que realiza un recuento del número de éxitos en un número de ensayos de Bernoulli independientes "n", y cuya probabilidad de que acontezca el éxito es igual a "p", quedando definida de la variable aleatoria de la forma $X \sim B(n,p)$. Por tanto, la relación existente entre ambas distribuciones la encontramos cuando el número de ensayos "n" es igual a uno. Al igual que realizamos en el caso de la Bernoulli, la función de probabilidad, formulación y principales propiedades de la distribución Binomial son:

- **Función de Probabilidad[4]:** $f(x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}$;
  $p \in [0,1]$ y $x = \{0,1,\dots,n\}$
- **Esperanza Matemática:** $E[X] = n \cdot p$
- **Varianza:** $Var[X] = n \cdot p \cdot (1-p) = n \cdot p \cdot q$
- **Función Generadora de Momentos:** $mgf = (q + pe^t)^n$

Otra de las características fundamentales que posee la distribución binomial se encuentra en que, para una probabilidad "p" igual a 0,5 y un número "n" elevado, se puede aproximar a una distribución normal, caracterizada por tener media, moda y mediana iguales al parámetro "μ" y simetría entorno a la media.

[4] Nótese que $\binom{n}{x} = \frac{n!}{x1 \cdot (n-x)!}$; y $n!$, $x!$ equivalen a los factoriales de "n" y "x" respectivamente

Fundamentos Teóricos de los Modelos de Regresión

Una vez conocida la distribución estadística que parece seguir la variable objeto de análisis, y antes de comenzar con la estimación de distintos modelos de regresión que nos permitan analizar el impacto, la significatividad y la calidad de las predicciones que estos ofrecen, se ha considerado necesario explicar los fundamentos de los modelos de regresión, desde los más sencillos como son los lineales con un único regresor que explique la variable de interés, hasta los modelos de regresión con variable dependiente binaria, haciendo mención en el camino a los de regresión múltiple y los modelos no lineales, siendo en los segundos en los que más nos detendremos.

Modelos de Regresión Lineal Simple

Buscan explicar el comportamiento de una variable, denominada dependiente, a través de una única variable explicativa. A modo de ejemplo, si buscásemos analizar la relación que tienen los capitales asegurados sobre la prima, en un gráfico de dispersión veríamos las distintas combinaciones (X, Y) para cada una de las observaciones de la muestra y estimaríamos un modelo que intentara capturar el impacto de "X" sobre la variable objetivo "Y".

Matemáticamente, esta relación quedaría expresada de la siguiente forma:

$$Y_i = \beta_0 + \beta_1 \cdot X_i + u_i \quad dónde:$$

- β_0: recibe el nombre de intercepto y corresponde con el término constante de la ecuación, pues no depende de las observaciones "i". Es el valor que tomaría la variable dependiente si el valor de "X" para la observación "i" fuese igual a cero. En ocasiones, este término no aporta ningún tipo de información útil para explicar "Y", por ello es conveniente comprobar su significatividad y la posible intuición económica detrás del resultado.
- β_1: denominada pendiente, determina el impacto que tiene sobre la variable dependiente "Y" el valor que la variable explicativa "X" ha tomado para cada "i". Cuanto mayor sea el valor de β_1 más grande será el impacto que tendrá sobre "Y" el valor que tome "X".

Los valores tomados por β_0 y β_1 reciben el nombre de coeficientes y representan los parámetros de la recta de regresión. Siendo la pendiente β_1 el cambio que experimenta la variable "Y" ante un aumento unitario de la variable "X". Estos coeficientes se calculan mediante Mínimos Cuadrados Ordinarios (MCO)

- u_i: denominado término de error del modelo, recoge el peso que tienen todos aquellos factores o variables que no han sido incluidos en la calibración del modelo pero que provocan las diferencias existentes entre los valores reales de las observaciones que componen la muestra y los valores que se habían estimado.

Modelos de Regresión Lineal Múltiple

Uno de los mayores problemas que existe dentro de la generación de modelos de regresión es lo que en estadística se denomina "sesgo por variable omitida" que corresponde a aquellos factores que tienen un impacto significativo sobre la variable dependiente pero que, al no incluirse dentro del modelo mediante otro regresor, acaba siendo absorbido por el coeficiente β_1 y por el término de error del modelo, lo que desemboca en modelos de reducida capacidad predictiva y, generalmente, con coeficientes sobreestimados.

La alternativa a tal efecto comienza en los modelos de regresión múltiple, que introducen para la estimación de coeficientes nuevos regresores que capturen los efectos que las distintas variables explicativas tienen sobre la variable de interés. Al introducir un mayor número de factores que aporten información sobre "Y", conseguiremos mejorar la capacidad para elaborar predicciones por el modelo. Si es cierto que, en este punto, se deben realizar test de multicolinealidad para evitar que dos o más variables incluidas como factores estén explicando un mismo efecto; por ejemplo: en la base de datos empleada se tiene información de si la póliza es de nueva producción (antigüedad en cartera menor o igual a un año) y otra sobre la antigüedad de las pólizas, agrupadas en función de si llevan en cartera un año o menos, entre uno y cinco o más de cinco; si incluyésemos ambas variables, se estaría introduciendo el efecto de que lleve en cartera un año o menos dos veces, por lo que finalmente emplearemos la que clasifica la antigüedad en tres grupos, ya que aporta más información.

Estos modelos, por tanto, buscan analizar el comportamiento de "Y" en función de un conjunto de variables explicativas "X_j". Matemáticamente se expresaría con la siguiente ecuación lineal:

$$Y_i = \beta_0 + \beta_1 \cdot X_{1,i} + \beta_2 \cdot X_{2,i} + \cdots + \beta_k \cdot X_{k,i} + u_i \quad dónde:$$

- β_0: representa al igual que en el de regresión simple, el intercepto del modelo.
- β_j: representa el coeficiente, impacto o pendiente de cada una de las variables explicativas $X = \{X_1, X_2, \dots, X_k\}$ que se han incluido en el modelo.

Un caso especial se da cuando alguna de las variables explicativas es binaria, es decir, cuando puede tomar únicamente los valores cero y uno. En este caso, el coeficiente β_j no representará una pendiente, sino la diferencia entre la media poblacional cuando la característica toma el valor uno, frente a cuando esta es cero.

Modelos No Lineales

Los modelos de regresión anteriores, a excepción de aquel con regresor binario, se caracterizan por tener como función de regresión poblacional una lineal, lo que implica que su pendiente sea constante. La realidad es que la relación existente entre las variables explicativas y la dependiente no siempre es lineal, tomando la siguiente metodología para su detección:

1. Identificación de una posible relación no lineal mediante el uso de la teoría económica y su representación gráfica univariante.
2. Especificación y estimación de una función no lineal mediante el método MCO.
3. Determinación de si el modelo no lineal mejora empíricamente los resultados obtenidos del modelo lineal, es decir, si los datos se ajustan de mejor forma.
4. Estimación del efecto de un cambio en "X" sobre "Y".

Dentro de este tipo de modelos diferenciamos tres grandes grupos:

- **Modelos Polinómicos:** en que la variable explicativa se introduce a través de una relación con sus potencias de grado "r". En el caso de la regresión simple:

$$Y_i = \beta_0 + \beta_1 \cdot X_{1,i} + \beta_2 \cdot X_{1,i}^2 + \dots + \beta_r \cdot X_{1,i}^r + u_i$$

- **Modelos con Interacciones:** surgen cuando cambios en la variable "X_{1i}" afectan tanto a la variable "Y_i" como a otra variable "X_{2i}". Son de gran utilidad para dividir la población en grupos que poseen diferentes medias. Si definimos "D_k" como variables dicotómicas "k":

$$Y_i = \beta_0 + \beta_1 \cdot D_{1,i} + \beta_2 \cdot D_{2,i} + \beta_3 \cdot (D_{1,i} \cdot D_{2,i}) + u_i$$

- **Modelos Logarítmicos:** convierten las variaciones que toman las variables en cambios porcentuales ya que muchas relaciones tienen una expresión natural en términos de porcentajes. Recordar que por teoría matemática, el logaritmo natural es la inversa de la función exponencial. Podemos distinguir tres tipos de relación:
    - Modelos Lineal – Log: dónde una variación del uno por ciento en "X_k" está asociada con un cambio en "Y" de $0{,}01 \cdot \beta_k$

$$Y_i = \beta_0 + \beta_1 \cdot \ln(X_{1,i}) + u_i$$

    - Modelos Log – Lineal: un cambio unitario en "X_1", es decir, $\Delta X_i = 1$ está asociado con un cambio de $100 \cdot \beta_i\%$

$$\ln(Y_i) = \beta_0 + \beta_1 \cdot X_{1,i} + u_i$$

    - Modelos Log – Log: es quizás el más importante de los tres ya que expresa como una variación del uno por ciento en "X_i" está asociada con una variación en "Y" de un $\beta_i\%$; por lo que β_k podría considerarse la elasticidad de "Y" respecto a "X_k"

$$\ln(Y_i) = \beta_0 + \beta_1 \cdot \ln(X_{1,i}) + u_i$$

Modelos con Variable Dependiente Binaria

Una vez realizado un breve repaso de los modelos de regresión que generalmente se emplean para el estudio de variables continuas, junto con las posibles relaciones no lineales que puedan existir entre sus variables explicativas, resulta ineludible la búsqueda de un modelo que se ajuste correctamente a la forma en que se presenta la variable de análisis, que recordemos es de carácter binario, tomando el valor uno en caso de impago y cero en caso contrario.

Modelo de Probabilidad Lineal

A tal efecto, se han desarrollado modelos de probabilidad lineal, basados en el ajuste de un modelo de regresión múltiple que tenga en cuenta el comportamiento binario de la variable dependiente. Por ello, esta función de regresión a través de sus coeficientes mostrará la probabilidad de que la variable objetivo tome el valor uno, dado una serie de factores "$X_{k,i}$". La ecuación (1) recoge la expresión del modelo de regresión lineal múltiple pero que, dado que "Y" es binaria, a través de la igualdad reflejada en la ecuación (2), podría expresarse en términos de probabilidad (3):

$$Y_i = \beta_0 + \beta_1 \cdot X_{1,i} + \beta_2 \cdot X_{2,i} + \cdots + \beta_k \cdot X_{k,i} + u_i \qquad (1)$$

$$E[Y|X_1, \dots, X_K] = \Pr(Y = 1|\, X_1, \dots, X_K) \qquad (2)$$

$$\Pr(Y = 1|X_1, \dots, X_K) = \beta_0 + \beta_1 \cdot X_1 + \cdots + \beta_k \cdot X_k \qquad (3)$$

Destacar que, en los modelos de regresión múltiple, cuando tratamos los coeficientes β_k como el impacto de la variable "k" sobre la dependiente "Y", se refiere manteniendo el resto de los factores constantes (ceteris paribus).

Modelos Lineales Generalizados

Los modelos lineales generalizados (GLM), como su propio nombre indica, son una generalización de los modelos de regresión lineal que previamente se han mencionado y cuyo objetivo es relacionar la aleatoriedad en que se enmarca la distribución de la variable dependiente con el componente no aleatorio mediante la utilización de la denominada "función link o de enlace".

Dentro de los modelos lineales generalizados se pueden diferenciar tres elementos esenciales:

- Componente aleatorio: se corresponde con la variable aleatoria "Y" y la función de distribución de probabilidad que esta sigue, dados los valores observados en la muestra $\{y_1, y_2, \dots, y_n\}$.

  En este tipo de modelos, la distribución de la variable "Y" pertenece a la familia exponencial (normal, gamma, binomial, poisson, …), lo que las hace realmente interesantes por sus propiedades dentro del campo de la estadística y la práctica actuarial, destacando:

  - La esperanza matemática de "Y" es igual a la media "μ", que depende de la relación existente entre las variables explicativas "X" introducidas en el modelo:

    $$\mathbb{E}[Y] = \mu = g^{-1}(X\beta) \quad \begin{cases} \mathbb{E}[Y] \equiv Valor\ Esperado\ de\ \mathrm{Y} \\ X\beta \equiv Combinación\ Lineal\ de\ Variables\ Explicativas \\ g \equiv función\ link \end{cases}$$

  - Permite la obtención de soluciones cerradas para los estimadores de máxima verosimilitud.
  - Posibilita la estimación bayesiana mediante el empleo de distribuciones conjugadas.

- Componente sistemático: se encarga de especificar cuáles son las variables explicativas que se encuentran incluidas en el modelo lineal, en forma de efectos fijos:

$$\alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

La expresión anterior es una combinación lineal compuesta por las variables explicativas, y que pueden ser recogidas en forma vectorial: $(\eta_1, \eta_2, \dots, \eta_k)$, de forma que, siendo $X_{i,j}$el valor que toma para la observación "i" la variable predictora "j":

$$\eta_i = \sum_j \beta_j X_{i,j}$$

Estos modelos permiten la introducción como regresores representaciones no lineales de sus variables, bien tengan como origen la interacción de varias variables ($X_4 = X_1 \cdot X_3$) o una expresión polinómica de alguna de ellas ($X_5 = X_2 + X_2^2$).

- Función de enlace: relaciona el valor esperado de la variable aleatoria "Y", que habíamos definido como $\mathbb{E}[Y] = \mu$, con su predictor lineal de la siguiente forma:

$$g(\mu) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

De esta forma, se consigue relacionar los componentes previamente mencionados para cada una de las observaciones $i = \{1, 2, \dots, N\}$. Matemáticamente:

$$\mu_i = \mathbb{E}[Y_i]$$

$$\eta_i = g(\mu_i) = \sum_j \beta_j X_{i,j}$$

Cabe destacar cómo los modelos GLM son una generalización, valga la redundancia, de los modelos lineales clásicos, siendo el caso dónde la función de enlace $g(\mu)$ es igual al valor esperado "μ", conocido también como función identidad:

$$E(Y) = \mu = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Si bien los modelos lineales generalizados permiten el análisis de variables continuas, como puede ser la severidad (coste de los siniestros) mediante la utilización de distintas

distribuciones además de la normal, la potencia es aún mayor si tenemos en cuenta que permite también el análisis de variables binarias, como se detalla a continuación.

En numerosas ocasiones, la variable dependiente puede tomar únicamente dos valores o categorías, es decir, es binaria; implicando que la variable aleatoria "Y" se distribuya como una binomial: $Y \sim Bin(1, \pi)$, dónde $y = \{0,1\}$

$$f(y|\pi) = \pi^y \cdot (1-\pi)^{1-y} = (1-\pi) \cdot \left(\frac{\pi}{1-\pi}\right)^y = (1-\pi) \cdot \exp\left[ylog\left(\frac{\pi}{1-\pi}\right)\right]$$

Obteniendo el parámetro natural de "Y" (4) que implica que la esperanza de "Y" sea (5), que dependerá de las "k" variables explicativas que se hayan considerado incluir en el análisis y que conforman el vector $X = (X_1, X_2, \dots, X_k)$, siendo su varianza la propia de una binomial (6):

$$Q(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = logit(\pi) \qquad (4)$$

$$E(Y) = P(Y = 1) = \pi(X) \qquad (5)$$

$$Var(Y) = \pi(X) \cdot (1 - \ \pi(X)) \qquad (6)$$

De esta forma, suponiendo la relación lineal de las variables, obtendríamos un modelo similar al de regresión lineal pero adaptado a respuestas binarias, que se puede definir como un modelo de probabilidad lineal, ya que expresa la variación de la probabilidad de ocurrencia del evento definido como éxito de forma lineal respecto a la variable dependiente "X".

$$\pi(X) = \alpha + \beta \cdot X$$

Siendo el coeficiente β el cambio que provoca un cambio unitario de "X" sobre la probabilidad de éxito.

El principal problema que muestran estos modelos de probabilidad se encuentra en que tratan la variable como si fuese lineal, provocando que los resultados que estos pueden exportar tomen valores superiores a uno o inferiores a cero, cuando es conocido que la probabilidad sólo puede tomar valores comprendidos entre cero y uno, ambos incluidos. Para dar solución a este problema, se han desarrollado los modelos no lineales "probit" y "logit", específicos para el análisis de variables dependientes binarias.

Modelo Probit

El modelo de regresión probit utiliza la función de distribución de probabilidad acumulada correspondiente a la normal estandarizada sobre la combinación de variables que se han incluido, permitiendo así que los valores obtenidos expresen una probabilidad acotada entre cero y uno. En términos matemáticos, la relación existente quedaría descrita de la siguiente forma:

$$\pi(X) = F(X) = \Phi(X)$$

Por tanto, si atendemos a que la función de X corresponde con el conjunto de variables explicativas que queremos introducir en nuestro modelo, en el caso más simple de existencia de un único regresor, quedaría expresado como:

$$\pi(X) = \Phi(\alpha + \beta \cdot X)$$

$$\Phi^{-1}\big(\pi(X)\big) = \alpha + \beta \cdot X$$

La forma más sencilla de entender el proceso es pensar en el término $\alpha + \beta \cdot X$ como el resultado de aplicar los coeficientes sobre el valor de la observación para dicha variable e interpretarlo como el "z" que buscaremos en la tabla de la distribución normal acumulada estandarizada; a modo de ejemplo, si el término aportara un valor de 1,5 la probabilidad asociada sería igual a $0{,}066807 = 6{,}68\%$.

Para el caso de la regresión múltiple, esencial dado el importante sesgo por variable omitida que tienen los modelos de regresión simple, la manera de proceder es similar salvo porque en este nuevo caso el número de regresores es mayor. Su ecuación se describiría como se refleja a continuación:

$$\Pr(Y = 1|X_1, X_2, \dots, X_k) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)$$

Con este modelo, el proceso de cálculo de probabilidades resulta relativamente sencillo, pues una vez obtenidos los coeficientes "β_k" mediante el método de máxima verosimilitud y tras incorporar los valores observados para cada una de las variables incluidas, únicamente faltaría por buscar en las tablas de la distribución normal estándar la probabilidad que este tiene asociada, es decir, el valor "z" a buscar sería:

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Los coeficientes se interpretan como el cambio en el valor de "z" asociado a un cambio unitario de una de las variables dependientes, dejando constantes las demás; en tal caso, sería necesario calcular la probabilidad existente antes de la alteración de "X" y la probabilidad asociada al nuevo valor de "z". Por ello, se precisa de la transformación mediante la función link de las predicciones del modelo para poder traducir los resultados a probabilidades y dotarles de una interpretación económica. Además, una ventaja que presenta la estimación de los coeficientes por máxima verosimilitud es su consistencia y que se distribuyen normalmente para muestras grandes, lo que permite construir intervalos de confianza y calcular los estadísticos t.

Modelo Logit

Frente al modelo probit, mencionado en líneas anteriores, aparecen los modelos logit como otra forma de modelizar relaciones no lineales con variables dependientes binarias. Matemáticamente, el modelo logit con varios regresores se expresaría con la ecuación (7), con función de enlace (8):

$$\Pr(Y = 1|X_1, X_2, \dots, X_k) = \mathrm{F}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k) \quad (7)$$

$$\pi(X) = F(X) = \frac{e^{\boldsymbol{X}}}{1 + e^{\boldsymbol{X}}} \quad (8)$$

$$1 - \pi(X) = 1 - \frac{e^{\boldsymbol{X}}}{1 + e^{\boldsymbol{X}}} = \frac{1}{1 + e^{\boldsymbol{X}}} \rightarrow \frac{\pi(X)}{1 - \pi(X)} = e^{\boldsymbol{X}} \rightarrow \log\left(\frac{\pi(X)}{1 - \pi(X)}\right) = \boldsymbol{X}$$

Entendiendo como "X" la función que conforma el componente sistemático del modelo, es decir $\boldsymbol{X} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$. Por tanto, la probabilidad de éxito del suceso sería la obtenida mediante la ecuación (9):

$$\Pr(Y = 1|X_1, X_2, \dots, X_k) = \mathrm{F}\frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}} \quad (9)$$

La principal diferencia de estos respecto a la modelización probit, es el uso como función de distribución la logística, que se define en términos de la función exponencial, y no de la distribución normal estandarizada; si bien los coeficientes también se pueden estimar por máxima verosimilitud, siendo este consistente y normalmente distribuido para muestras grandes.

Modelos de Variables Instrumentales

Tomando como base el modelo de regresión que relaciona una variable dependiente "Y" con un conjunto de variables independientes "X_j", y en el que el término de error "u_i" representa los factores omitidos que influyen en "Y", es conocido que el estimador de mínimos cuadrados ordinarios, con el que generalmente se desarrollan estos, presentan coeficientes inconsistentes si las variables independientes están correlacionadas con el término de error. Este fenómeno se produce por el sesgo de variable omitida, dónde factores que explican la variable dependiente no son incluidos como regresor y están relacionadas con alguna variable explicativa.

En consecuencia, surgen los modelos de variables instrumentales, diferenciando dos tipos de variables explicativas: endógenas en caso de estar correlacionadas con el término de error y exógenas en caso contrario. Para corregir el problema de inconsistencia de los estimadores de las variables endógenas, se emplearán una o varias variables denominadas instrumentos "Z", que deberán cumplir las siguientes condiciones:

- Relevancia: la variación de "X_i" está relacionada con la variación del instrumento "Z_i" → $corr(Z_i, X_i) \neq 0$
- Exogeneidad: el instrumento "Z_i" no está relacionado con el error del modelo estimado "u_i" → $corr(Z_i, X_i) \neq 0$

Por tanto, un instrumento que sea exógeno y relevante captura las variaciones de "X_i" que son exógenas, permitiendo así la consistencia de los coeficientes estimados por el modelo.

Para la correcta estimación de los parámetros en modelos en los que el instrumento cumple las condiciones de exogeneidad y relevancia, surge el estimador de mínimos cuadrados en dos etapas:

- Etapa 1: regresión que relaciona la variable "X" con el instrumento "Z".

$$X_i = \pi_0 + \pi_1 \cdot Z_i + v_i$$

- Etapa 2: regresión que relaciona la variable "Y" con la variable "X".

$$Y_i = \beta_0 + \beta_1 \cdot X_i + u_i$$

Con este procedimiento, obtendríamos los estimadores β_0 y β_1 por mínimos cuadrados en dos etapas, mediante el cálculo previo de los estimadores π_0 y π_1 por mínimos cuadrados ordinarios, todos ellos consistentes al no estar correlacionados con los errores del modelo.

Metodología Empleada

Una vez presentados algunos de los modelos de regresión existentes y sus principales diferencias, atendiendo a las características de la muestra de datos objeto de análisis, en que la variable dependiente es binaria, resulta evidente que debemos emplear modelos que capturen este hecho; siendo los modelos probit, logit y el de probabilidad lineal los que comparten este atributo.

Antes del desarrollo de softwares informáticos que introdujeran los algoritmos necesarios para el cálculo de los coeficientes de los modelos probit y logit, era el de probabilidad lineal el que estaba más extendido por su simplicidad de cálculo; sin embargo, actualmente son los dos primeros los más utilizados ya que su capacidad predictiva es notablemente superior, al capturar el comportamiento no lineal de las variables explicativas que los conforman. Además, el modelo de probabilidad lineal presenta un notable problema: puede reportar valores superiores a uno e inferiores a cero, lo cual no tendría sentido al ser la probabilidad mínima de impago igual a cero (0%) y la máxima igual a 1 (100%).

Por ello, dada la no linealidad de los datos, se ha considerado utilizar los modelos probit y logit para la estimación de los coeficientes por máxima verosimilitud, dado que puede aportar resultados consistentes y normalmente distribuidos, dado el gran tamaño de la muestra, y que permitan realizar predicciones de la probabilidad de impago de prima con valores comprendidos entre cero y uno. Destacar que los distintos modelos a estimar se realizarán bajo ambas metodologías, ya que los resultados aportados por estas se comportan, a priori, de forma muy similar; si bien compararemos cuál de ellos se ajusta mejor a los datos, en base a su capacidad predictiva.

A tal efecto, se hará uso del paquete estadístico “R Estudio”, realizando una división de la muestra global en dos submuestras, la primera compuesta por setenta y cinco mil observaciones, que se empleará para la estimación de los modelos, y la segunda formada por casi treinta mil, que irá dedicada a la validación de estos mediante el análisis de sus predicciones.

Estimación de Modelos – Predicción y Evaluación

Puesto que ya se ha definido cuál es la técnica que se va a seguir, centrada en los modelos lineales generalizados logit y probit y con el objetivo de estructurar el análisis de la forma más adecuada posible, se seguirá el siguiente proceso:

1. Estimación de modelos que incluyan un elevado número de variables explicativas, que se presentan como datos propios de la póliza y de las que se tiene información para la totalidad de la cartera.
2. Proceso de depuración de los modelos generados anteriormente, para estimar otros nuevos que sólo incluyan aquellas variables que resulten significativas en la explicación de la probabilidad de impago.
3. Elaboración de predicciones con esta primera serie de modelos, evaluación de estas para distintos niveles de probabilidad y análisis de la matriz de confusión asociada. Además se generará un rating crediticio en función de las probabilidades de impago estimadas.
4. Introducción de variables macroeconómicas para completar los modelos previamente estimados, analizando su impacto sobre la variable de interés.
5. Elaboración y análisis de las predicciones resultantes con los nuevos modelos, siguiendo el mismo criterio que el empleado en el punto 3.
6. Comparación y estudio de las mejoras encontradas con la introducción de los indicadores macroeconómicos mencionados.
7. Presentación del modelo final.

Estimación de Modelos I

Mencionar que, para simplificar la forma en que las variables son transformadas en ecuaciones, se ha definido la variable dependiente, es decir, la situación de impago/pago como "$Y1$" y las variables explicativas como "X_j" [5].

El primero de los modelos que se ha estimado busca la relación entre la probabilidad de impago de prima con el sector económico al que se dedica la empresa objeto de seguro, si su antigüedad es igual o inferior al año, el mes en que fue contratada la póliza, la tenencia o no de coaseguro y reaseguro, el capital y la prima neta, la periodicidad de pago (en términos anual, semestral, trimestral o irregular), el número acumulado de siniestros, la posible domiciliación bancaria del pago, el tramo y canal por el que entró

[5] La relación de variables explicativas y su identificador "X_j" se encuentran recogidas en el anexo 1 del documento.

la prima, su localización geográfica y el tipo de mediador de la póliza. Dada la extensión de este modelo, se encuentra recogido en el anexo 2.1 (modelo 1).

Analizando los resultados obtenidos, se ha encontrado que en ambas modelizaciones (probit y logit) las variables que aparecen como significativas coinciden. Además, gracias a la potencia del paquete estadístico, se puede observar el coeficiente y nivel de significatividad asociado a cada uno de los posibles valores que toman las variables de tipo factor que se encuentran incluidas. A tal efecto, encontramos como dentro del sector económico, aquellas pólizas dedicadas a la prestación de servicios (X4S3) presentan probabilidades de impago significativamente diferentes respecto a las que se enmarcan en los sectores primario y secundario. De igual modo, cuestiones como el capital y la prima neta, la existencia de reaseguro, el tipo de canal por el que fue emitida la póliza y su zona geográfica, el método en que se paga la prima (tanto en periodicidad como en su posible domiciliación), la categoría del mediador o el número acumulado de siniestros también presentan una gran significatividad.

Cabe resaltar también lo acontecido con el mes en que se originó la póliza, ya que observamos cómo algunos meses si ayudan a explicar el impago de la prima. Por ello, y apoyándonos en la gráfica mostrada en el análisis descriptivo de las variables, se introducirá en los modelos posteriores una variable que muestre el trimestre en que se originó en sustitución de los meses. Sorprende también la no significatividad que parece tener el que la póliza sea o no de nueva producción; sin embargo, para poder garantizar que la antigüedad no es relevante en la probabilidad de impago, emplearemos una nueva variable que categoriza las pólizas en función de si su antigüedad es igual o menor al año, entre uno y cinco o si lleva más de cinco años en cartera.

Además, dado que el capital y la prima neta pueden generar problemas de multicolinealidad de variables, es decir, que ambas expliquen un mismo efecto por motivo de su correlación, se ha considerado su cálculo, obteniendo como resultado un porcentaje inferior al cincuenta por ciento, por lo que no se produce un efecto de multicolinealidad entre ellas.

	Capital	***Prima Neta***
Capital	1	
Prima Neta	0,4360	1

De esta forma, se han estimado los dos nuevos modelos (modelo 2) que se muestran a continuación, que incluirán únicamente aquellas variables que se presentaban como significativas en los anteriores, junto con la nueva caracterización de la antigüedad en cartera, mencionada en líneas previas.

Variables	Logit			Probit		
	Coeficiente	Std. Error		Coeficiente	Std. Error	
Intercepto	-1,869	0,115	***	-1,110	0,061	***
X4S2	-0,005	0,044		0,006	0,024	
X4S3	0,022	0,025		0,020	0,014	
X9	0,097	0,006	***	0,055	0,003	***
X121	-0,605	0,061	***	-0,321	0,030	***
X14	-0,179	0,015	***	-0,100	0,008	***
X202	-0,392	0,025	***	-0,220	0,014	***
X203	-0,606	0,030	***	-0,331	0,017	***
X22BR	-0,260	0,086	**	-0,132	0,045	**
X22C	-0,100	0,024	***	-0,056	0,013	***
X22D	0,293	0,436		0,166	0,250	
X22O	-0,301	0,055	***	-0,138	0,030	***
X2310	0,297	0,044	***	0,162	0,024	***
X232	0,068	0,039	.	0,039	0,021	.
X233	0,190	0,038	***	0,108	0,020	***
X234	0,282	0,035	***	0,155	0,019	***
X235	0,487	0,037	***	0,270	0,021	***
X239	-1,243	0,128	***	-0,599	0,060	***
X27I	-2,227	0,454	***	-1,019	0,179	***
X27S	0,497	0,025	***	0,272	0,014	***
X27T	0,561	0,032	***	0,295	0,018	***
X291	0,237	0,035	***	0,123	0,019	***
X331	-0,362	0,025	***	-0,197	0,014	***
X72	-0,002	0,003		0,000	0,001	
X962	0,110	0,029	***	0,060	0,016	***
X963	0,056	0,030	.	0,028	0,017	.
X964	-0,007	0,029		-0,006	0,016	
Significatividad: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

En este caso, el comportamiento mostrado por ambos modelos probit y logit sigue coincidiendo en cuanto a variables significativas se refiere, destacando la antigüedad entre ellas, ya que previamente parecía no ser un factor relevante en la explicación de la probabilidad de impago, frente a lo que la intuición económica sugiere de que clientes con más tiempo operando con la entidad será más probable que paguen la prima acordada. Sin embargo, el número acumulado de siniestros que había sufrido la entidad asegurada en años previos y el sector de actividad en que opera la empresa dejan de ser significativos, siendo este último algo destacable, pues cabría pensar que en función de la actividad desarrollada, las probabilidades de impago fueran diferentes.

Pese a ello, realizaremos dos modelos paralelos, recogidos en el anexo 2.2, uno que mantenga en el análisis este último concepto y otro que lo desestime, comprobando así

la posible necesidad de controlar la actividad empresarial con la comparación de las predicciones que estos reportan.

A raíz de las estimaciones obtenidas para los coeficientes, los efectos de incluir el sector de actividad parecen mínimos, si bien será necesario analizar la calidad de las predicciones sobre la muestra de validación para comprobar si, al no distar mucho los resultados obtenidos con su inclusión u omisión, es preferible no tomarla en consideración. En cuanto a las variables incluidas, se ha encontrado un impacto significativo de todas ellas, si bien para determinados valores de algunos factores como la emisión por negocio directo, frente a otros como bróker o corredores, no tiene una influencia estadística relevante en la explicación del impago de la prima. Lo que si resulta evidente es la diferencia de comportamientos por zona geográfica y la importancia que tiene el método de pago en este hecho.

Dado que desgranar la forma de pago en función de su periodicidad puede reducir la repercusión del pago fraccionado frente al pago único, en términos netos, se han estimado los modelos anteriores modificando la variable mencionada, de forma que tenga únicamente en cuenta este concepto de forma binaria y no su grado de fraccionamiento. A continuación se muestra el resultado obtenido para el segundo de los modelos anteriores (modelo 9), si bien en el anexo 2.3 se encuentra también el estimado para el tercero de los modelos previos (modelo 10).

Variables	Logit			Probit		
	Coeficiente	Std. Error		Coeficiente	Std. Error	
Intercepto	-1,474	0,118	***	-0,887	0,063	***
X4S2	-0,003	0,044		0,007	0,024	
X4S3	0,021	0,025		0,019	0,014	
X9	0,097	0,006	***	0,055	0,003	***
X121	-0,604	0,061	***	-0,320	0,030	***
X14	-0,170	0,014	***	-0,096	0,008	***
X202	-0,374	0,025	***	-0,210	0,014	***
X203	-0,588	0,030	***	-0,320	0,017	***
X22BR	-0,249	0,086	**	-0,128	0,045	**
X22C	-0,097	0,024	***	-0,055	0,013	***
X22D	0,301	0,435		0,170	0,250	
X22O	-0,286	0,055	***	-0,133	0,030	***
X2310	0,298	0,044	***	0,162	0,024	***
X232	0,065	0,039	.	0,038	0,021	.
X233	0,188	0,038	***	0,106	0,020	***
X234	0,281	0,035	***	0,155	0,019	***
X235	0,486	0,037	***	0,270	0,020	***
X239	-1,242	0,127	***	-0,607	0,059	***
X28U	-0,501	0,022	***	-0,271	0,012	***
X291	0,263	0,035	***	0,136	0,018	***
X331	-0,360	0,025	***	-0,196	0,014	***
X72	-0,003	0,003		-0,001	0,001	
X962	0,108	0,029	***	0,059	0,016	***
X963	0,052	0,030	.	0,026	0,017	
X964	-0,012	0,029		-0,009	0,016	

Una vez obtenido este conjunto de modelos, sobre los que se ha llevado a cabo el análisis de coeficientes, y continuando con el guion previamente establecido, resulta necesario la realización de predicciones sobre la muestra que se había reservado para su validación. Por consiguiente, y dadas las facilidades con la que estas se pueden realizar con los softwares actualmente disponibles, se llevarán a cabo para la totalidad de modelos estimados los procesos de predicción de impago de cada una de las observaciones.

Dado que los modelos van a definir una probabilidad de que el tomador no pague la prima, se establecerá una probabilidad a partir de la cual se considerará que la póliza impaga, y que será similar para la totalidad de modelos a evaluar, ya que sino la valoración estaría sesgada y no sería correcta. A tal efecto, se ha realizado un doble análisis en el cálculo de las tasas de impago:

1. Haciendo uso de distintos niveles de probabilidad a partir de los que consideraríamos que la póliza va a incurrir en el impago de la prima, se ha calculado el número de casos en que el modelo no fue capaz de predecirlo.
   Además, para profundizar en su análisis, se han calculado las matrices de confusión para la totalidad de niveles de probabilidad contemplados, si bien se centrará el estudio en una probabilidad del treinta por ciento a partir de la que considerar el impago.

2. Se ha generado un rating crediticio que dota de una calificación, en función de las probabilidades de impago estimadas, para calcular el número de casos en que las pólizas de cada categoría no pagaron la prima. La forma en que este se ha establecido pretende generar una valoración similar a la que entidades de rating desarrollan para empresas y administraciones públicas.

Para el primero de los análisis es importante tomar en consideración que, al no tener en cuenta uno de los principales factores que motivan el riesgo de crédito, como es la situación financiera de los tomadores, los niveles de probabilidad a partir de los cuales consideraremos que una póliza no pagará la prima deberán ser inferiores al cincuenta por ciento que, generalmente, se toma para esta situación binaria.

A continuación se muestran los resultados obtenidos mediante la totalidad de modelos logit y probit previamente estimados, para los distintos niveles en los que consideraremos que una póliza incurrirá en impago de la prima:

	% de Impagos No Previstos por el Modelo											
	Modelo 1		Modelo 2		Modelo 3		Modelo 4		Modelo 9		Modelo 10	
	Logit	Probit	Logit	Probit	Logit	Probit	Logit	Probit	Logit	Probit	Logit	Probit
40%	99,9%	99,9%	99,5%	99,8%	99,5%	99,8%	99,5%	99,9%	99,6%	99,9%	99,6%	99,9%
30%	94,3%	95,3%	89,9%	90,9%	89,8%	90,9%º	89,9%	91,0%	90,6%	91,6%	90,6%	91,6%
20%	61,4%	61,2%	59,2%	58,8%	59,3%	58,8%	59,4%	58,7%	59,7%	59,3%	59,7%	59,4%
15%	34,8%	33,7%	35,9%	35,1%	35,9%	35,1%	36,1%	35,0%	36,1%	35,1%	36,1%	35,1%
10%	9,8%	9,6%	12,5%	11,9%	12,5%	11,9%	12,4%	12,0%	12,2%	11,7%	12,3%	11,7%
5%	1,7%	1,9%	1,7%	1,9%	1,7%	1,9%	1,7%	1,8%	1,7%	1,7%	1,7%	1,7%

Como se puede observar en la tabla anterior, los casos de impago reales que habían sido previstos por el modelo están enormemente influenciados por el porcentaje a partir del cual se considera que una póliza incurrirá en impago. De forma que, si se estableciese que sólo las pólizas con un porcentaje superior al cuarenta por ciento incurrirán en este, la calidad del modelo sería nula, puesto que no serviría para predecir este fenómeno. Sin embargo, y en línea de lo mencionado anteriormente, a medida que aumentamos la exigencia de este, en términos de probabilidades a partir de las que se considera el impago, los resultados van progresando de forma paulatina.

De igual modo, atendiendo a los resultados aportados por los distintos modelos, es el último de los anteriores (modelo 10) el que posee una mayor capacidad predictiva, destacando el criterio de estimación probit como el mejor de los empleados; cabe recordar como este fue el generado mediante la introducción del trimestre de origen de la póliza, los capitales asegurados y su prima, el sector de actividad que desarrolla, su canal de emisión, la categoría del mediador, la antigüedad en cartera, la zona geográfica en que se generó, si está domiciliado su pago, la tenencia de reaseguro y si el pago de la prima es único o fraccionado.

Para el modelo seleccionado, se ha realizado el null deviance test[6], cuyo objetivo es probar si la introducción de las variables consideradas logra explicar mejor el comportamiento que uno nulo. El resultado obtenido confirma que es acertada la elección del modelo, bajo el criterio del test propuesto.

[6] https://stats.stackexchange.com/questions/108995/interpreting-residual-and-null-deviance-in-glm-r

Además, haciendo uso de su matriz de confusión, se han calculado algunos ratios o indicadores, que ayudarán a realizar el conocido análisis ROC, que mide la sensibilidad respecto a su especificidad en un sistema de clasificación binario, ante variaciones en el umbral de discriminación empleado. A través de este, se podrá evaluar de forma más precisa la calidad del modelo, atendiendo a su nivel de aciertos y fallos, tanto para los casos de pago como de impago. Esta información, definiendo el caso positivo como el impago, aparece recogida en la siguiente tabla; dónde los verdaderos positivos (VP) y verdaderos negativos (VN) expresarían los impagos y pagos correctamente estimados, y los falsos positivos (FP) y falsos negativos (FN) que conforman los errores de predicción de tipo uno y dos, es decir, los casos en que se estimó que incurriría en impago y finalmente pagó, o los que se contempló su pago y no se produjo.

Modelo	Probit	**Valor Predicho**	
Probabilidad	30%	Pago (0)	Impago (1)
Valor Real	Pago (0)	24007	626
	Impago (1)	4059	371

$$\boldsymbol{VP} = 371$$
$$\boldsymbol{VN} = 24007$$
$$\boldsymbol{FP} = 626$$
$$\boldsymbol{FN} = 4059$$

Una vez recogidos los casos de pago e impago reales y estimados, se han calculado distintos ratios que aporten una visión más completa sobre el modelo y esta metodología de evaluación. Los resultados obtenidos para algunos de estos, que quedan definidos en el Anexo 5 del documento, reflejan una capacidad del 8.37% y del 97.45% para predecir los casos de impago (VPR) y pago (SPC) respectivamente; desembocando en una tasa de acierto (ACC) del 83.87%. Esto también se traduce en un nivel de falsos positivos (FPR) del 2.54, lo que prácticamente garantiza que las pólizas sobre las que se estimó un impago, finalmente no pagaron; siendo fundamental para poder tomar alguna medida previa. Además, los ratios predictivos positivos (PPV) y negativos (NPV) desprenden como el 37.21% de las pólizas que no iban a pagar al final no lo hicieron y el 85.53% de las que iban a pagar terminaron realizando el pago. Por último, se ha obtenido un 62.78% de ratio de falsos descubrimientos (FDR).

En referencia a la realización del rating crediticio, se han considerado las siguientes posibles calificaciones que una póliza puede alcanzar, en función de la probabilidad de impago estimada por el modelo:

Rating	Probabilidad Impago
A	$P(Y_i = 1) \in [0\,,0.03]$
B	$P(Y_i = 1) \in (0.03\,,0.075]$
C	$P(Y_i = 1) \in (0.075\,,0.125]$
D	$P(Y_i = 1) \in (0.125\,,0.25]$
E	$P(Y_i = 1) \in (0.25\,,0.40]$
F	$P(Y_i = 1) \in (0.40\,,1]$

Entendiendo las probabilidades de la tabla como las tasas poblacionales de impago de prima en una cartera de seguros de características similares a la analizada; por ejemplo, cabría esperar que de cada doscientas pólizas con rating C, entre quince y veinticinco no paguen la prima en el plazo previsto; o en este mismo sentido, de cada cien pólizas con rating F, más de cuarenta no pagarán la prima.

Con el criterio definido, se ha calculado para cada modelo el número de pólizas que quedaría enmarcado en cada rating y cuántas de ellas finalmente no pagaron la prima; para, finalmente, calcular las tasas de impago realmente acontecidas por calificación otorgada. Así, ha sido posible comparar el nivel de impagos que contemplaba cada rating con los que han terminado ocasionándose en cada grupo.

A continuación se presentan las tasas de impago que realmente se han producido, dado un determinado rating, para los modelos previamente comentados:

	Tasa de Impago Según Calificación											
	Modelo 1		**Modelo 2**		**Modelo 3**		**Modelo 4**		**Modelo 9**		**Modelo 10**	
Rating	Logit	Probit	Logit	Probit	Logit	Probit	Logit	Probit	Logit	Probit	Logit	Probit
A	3,5%	3,0%	4,3%	3,6%	4,3%	3,6%	4,3%	4,2%	4,1%	3,8%	4,1%	3,8%
B	4,7%	4,8%	5,2%	5,3%	5,2%	5,3%	5,1%	5,1%	5,3%	5,3%	5,3%	5,3%
C	10,3%	10,3%	10,8%	10,7%	10,8%	10,6%	10,8%	10,8%	10,7%	10,7%	10,7%	10,7%
D	17,8%	17,9%	17,4%	17,5%	17,4%	17,5%	17,4%	17,4%	17,4%	17,4%	17,4%	17,4%
E	29,9%	29,4%	31,5%	31,4%	31,4%	31,4%	31,4%	31,6%	31,4%	31,6%	31,2%	31,5%
F	3,1%	3,1%	41,4%	42,1%	42,1%	42,1%	39,3%	40,0%	35,6%	45,5%	37,0%	45,5%

Con este criterio, se aprecia de forma clara cómo los modelos que no funcionaban de forma precisa, al establecer una probabilidad fija a partir de la que considerar el impago, resultan tener una mayor capacidad predictiva si establecemos distintos niveles con los que se produce este fenómeno. De tal forma, y poniendo foco en el mejor de esta selección de modelos (modelo 10), aquellas pólizas que fueron dotadas con la

calificación A impagaron en el cuatro por ciento de los casos, frente a las B que lo hicieron en más del cinco por ciento de las ocasiones, las C en casi un once por ciento, las D en casi un diecisiete y las E en más de un treinta; y las calificadas como F en un 37% - 45.5% de los casos, en función de la metodología empleada en la modelización.

Si se comparan estos resultados, con los inicialmente contemplados en la elaboración del rating, se observa cómo las tasas de impago que realmente se han dado para cada grupo se aproximan a los rangos contemplados por los mismos; por lo que se puede afirmar que el criterio es válido. A modo de ejemplo, el criterio establecía que las pólizas cuya probabilidad de impago estimada se encontraran en el intervalo $(0.125\ ,\ 0.25]$ recibirían la calificación D, lo que supone que de cada doscientas, entre veinticinco y cincuenta incurrirían en impago; dado que en la muestra es el diecisiete por ciento de las pólizas las que experimentan este hecho, en los mismos términos, treinta y cinco de cada doscientas, se cumple la condición del intervalo descrito.

Posteriormente, han sido introducidos en el análisis algunos de los principales indicadores macroeconómicos que comúnmente son tomados en cuenta en la valoración de la situación económica de las regiones: la organización sectorial, medida a través del peso de los sectores primario, secundario y terciario como porcentaje del Producto Interior Bruto de la región; y las tasas de paro y de variación del paro que estas poseen. Con el objetivo de profundizar en el impacto de estos, en la explicación de la probabilidad de impago, se ha introducido la información por póliza y provincia, en función del último año en que estuvo en vigor.

Dado que se están introduciendo variables que pueden presentar una importante relación entre ellas, se ha llevado a cabo el cálculo de la siguiente matriz de correlaciones.

	Capital	***Prima Neta***	***Paro***	***Var_Paro***	***S1***	***S2***	***S3***
Capital	1						
Prima Neta	0,335	1					
Paro	-0,049	0,015	1				
Var_Paro	0,030	0,019	-0,166	1			
S1	-0,034	-0,035	0,414	-0,109	1		
S2	0,040	-0,002	**-0,591**	0,164	-0,195	1	
S3	0,062	0,004	-0,302	0,026	**-0,655**	-0,434	1

En ella advertimos como el paro y el peso del sector secundario en el PIB de la región albergan una correlación superior al cincuenta por ciento, posiblemente ocasionada por ser este sector el más representativo en términos de empleo estable, pues las actividades industriales precisan de una base de empleo alta; por lo que introduciremos en el modelo el sector secundario, ya que puede dotarnos de una mayor información acerca del tejido empresarial de la región en la que se ha producido el impago. Además, evitamos posibles problemas de multicolinealidad cruzada, dado que se incluye también la variación del paro. Lo mismo sucede con los sectores primario y terciario, por ello, se incluirán en la modelización sólo el terciario, que junto al secundario, aportarán una visión completa de la estructura económica de la región.

Como se comentó al inicio del epígrafe, sería el mejor de los modelos previamente calculados sobre el que introduciríamos los factores macroeconómicos para comprobar si la calidad de las predicciones de impago generadas mejora las inicialmente obtenidas. Pese a ello, en el anexo 2.4 se encuentran recogidos la totalidad de modelos estimados previamente pero añadiendo estas nuevas variables. Los resultados obtenidos para el modelo 10, tras la inclusión de la tasa de variación del paro y el porcentaje del PIB que representan los sectores secundario y terciario (modelo 18) son:

Variables	Logit			Probit		
	Coeficiente	Std. Error		Coeficiente	Std. Error	
Intercepto	3,988	0,203	***	2,207	0,113	***
X4S2	-0,005	0,045		0,007	0,024	
X4S3	0,040	0,025		0,032	0,014	*
X9	0,113	0,006	***	0,063	0,003	***
X121	-0,547	0,061	***	-0,286	0,030	***
X14	-0,193	0,014	***	-0,107	0,007	***
X202	-0,370	0,026	***	-0,207	0,015	***
X203	-0,585	0,030	***	-0,316	0,017	***
X22BR	-0,292	0,087	***	-0,146	0,045	**
X22C	-0,085	0,024	***	-0,049	0,013	***
X22D	0,260	0,441		0,143	0,252	
X22O	-0,404	0,056	***	-0,190	0,030	***
X2310	-0,049	0,046		-0,041	0,025	
X232	-0,061	0,040		-0,037	0,022	.
X233	-0,136	0,040	***	-0,079	0,022	***
X234	-0,166	0,041	***	-0,093	0,023	***
X235	-0,539	0,053	***	-0,306	0,030	***
X239	-1,509	0,130	***	-0,749	0,061	***
X28U	-0,478	0,023	***	-0,257	0,013	***
X291	0,288	0,035	***	0,146	0,019	***
X331	-0,349	0,025	***	-0,188	0,014	***
X962	0,103	0,029	***	0,055	0,016	***
X963	0,042	0,031		0,021	0,017	
X964	-0,009	0,030		-0,005	0,016	
X90	-0,169	0,582		-0,170	0,327	
X92	-8,285	0,302	***	-4,655	0,167	***
X93	-5,478	0,189	***	-3,125	0,107	***

Analizando la significatividad de los coeficientes, encontramos como el sector de actividad en que se engloba la entidad es estadísticamente relevante en la explicación de la probabilidad de impago con la modelización probit, frente a la logit en que no lo es. De igual forma, en las nuevas variables macroeconómicas incluidas encontramos dos efectos: en contra de lo pronosticado, la tasa de variación del paro no resulta significativa en la explicación del impago; posiblemente sea causa de ello el rango de años de valoración, en los que se ha producido un continuo crecimiento del empleo pero los casos de impago se han mantenido estables; indicando que pueda albergar un importante peso la conducta del tomador. Sin embargo, el peso de los sectores secundario y terciario en el PIB de la región si presentan una gran significatividad. Por ello, y previa validación con el estudio de las calidad de las predicciones que genera, parece necesaria su inclusión. A continuación se recogen las tasas de fallo de estos modelos, para los distintos niveles de probabilidad previamente empleados:

	% de Impagos No Previstos por el Modelo											
	Modelo 13		Modelo 14		Modelo 15		Modelo 16		Modelo 17		Modelo 18	
	Logit	Probit	Logit	Probit	Logit	Probit	Logit	Probit	Logit	Probit	Logit	Probit
40%	96,9%	97,6%	95,5%	96,4%	95,5%	96,3%	95,5%	96,3%	95,8%	96,5%	95,8%	96,5%
30%	87,1%	87,6%	83,4%	84,4%	83,5%	84,4%	83,7%	84,6%	84,1%	85,1%	84,1%	85,2%
20%	57,0%	56,8%	55,8%	55,1%	55,7%	55,2%	55,8%	55,2%	56,3%	55,4%	56,3%	55,4%
15%	34,4%	33,5%	34,5%	33,7%	34,5%	33,7%	34,4%	33,3%	34,6%	33,6%	34,6%	33,7%
10%	12,0%	11,6%	14,2%	13,8%	14,4%	13,8%	14,3%	13,6%	14,2%	13,5%	14,2%	13,5%
5%	2,0%	2,0%	2,1%	2,4%	2,2%	2,4%	2,1%	2,3%	2,1%	2,3%	2,1%	2,4%

En la tabla encontramos como la línea que seguían los modelos iniciales se mantiene, siendo los últimos (modelos 17 y 18) los que presentan menores tasas de fallo y cómo, a medida que reducimos el porcentaje exigido para la consideración de impago, estas se reducen de forma notable. Dado que no estamos considerando aspectos financieros de las pólizas y el número de variables consideradas es menor, en comparación con los sofisticados modelos de impago empleados por los bancos y otras entidades de crédito, el nivel de exigencia debería ser notablemente superior a la generalmente empleada del cincuenta por ciento; por lo que las probabilidades a partir de las que se definiría el impago deberían ser menores.

Realizando el null deviance test para el nuevo modelo se han extraído las mismas conclusiones, es decir, las variables consideradas logran explicar mejor el comportamiento que un modelo nulo.

La matriz de confusión de este nuevo modelo con variables macroeconómicas desprende los siguientes resultados:

Modelo	Probit	**Valor Predicho**		$\boldsymbol{VP} = 657$
Probabilidad	30%	Pago (0)	Impago (1)	$\boldsymbol{VN} = 23497$
Valor Real	Pago (0)	23497	1136	$\boldsymbol{FP} = 1136$
	Impago (1)	3773	657	$\boldsymbol{FN} = 3773$

En este caso, la capacidad de predicción de los casos de impago (VPR) aumenta hasta el 14.8% frente a la leve reducción que experimenta la asociada al pago (SPC), que se sitúa en el 95.38%, con una tasa de acierto (ACC) asociada del 83.11%; lo que implica que el porcentaje de falsos positivos (FPR) se sitúe en el 4.61%. Además, los ratios predictivos positivos (PPV) y negativos (NPV) se sitúan en el 36.64% y 86.16% respectivamente, con un ratio de falsos descubrimientos (FDR) del 63.35%.

De igual modo, y dado que uno de los objetivos que se habían fijado en este estudio era analizar el impacto que tiene la introducción de variables macroeconómicas en la calidad de las predicciones de los modelos iniciales, se han calculado las diferencias existentes entre las tasas de fallo de los modelos que las incluyen frente a los primeros, recogidas en la siguiente tabla:

	Diferencia por Inclusión de Variables Macroeconómicas											
	Modelo 1 vs 13		**Modelo 2 vs 14**		**Modelo 3 vs 15**		**Modelo 4 vs 16**		**Modelo 9 vs 17**		**Modelo 10 vs 18**	
	Logit	Probit	Logit	Probit	Logit	Probit	Logit	Probit	Logit	Probit	Logit	Probit
40%	3,0%	2,3%	3,9%	3,4%	4,0%	3,5%	4,0%	3,6%	3,8%	3,3%	3,8%	3,4%
30%	7,3%	7,8%	6,4%	6,5%	6,3%	6,5%	6,2%	6,4%	6,4%	6,5%	6,5%	6,5%
20%	4,4%	4,3%	3,5%	3,7%	3,5%	3,6%	3,6%	3,5%	3,4%	3,9%	3,4%	3,9%
15%	0,4%	0,2%	1,4%	1,4%	1,4%	1,4%	1,7%	1,7%	1,5%	1,4%	1,5%	1,5%
10%	-2,2%	-2,0%	-1,8%	-1,8%	-1,9%	-1,9%	-1,9%	-1,6%	-2,0%	-1,8%	-1,9%	-1,8%
5%	-0,3%	-0,2%	-0,4%	-0,6%	-0,5%	-0,5%	-0,5%	-0,5%	-0,4%	-0,6%	-0,4%	-0,7%

Como se puede observar, con el cálculo de las diferencias existentes entre los fallos de predicción de los modelos iniciales, frente a los que incluyen variables macroeconómicas, las mejoras son más que evidentes para la totalidad de modelos si bien, en función de la probabilidad a partir de la cual se considerará que la póliza no pagará la prima, se ha encontrado:

- Para probabilidades superiores al 15%, las tasas de error de predicción de impago se reducen, siendo a partir del 20% cuando lo hacen con mayor significatividad, siendo interesante la introducción de este tipo de variables.
- Cuando los niveles exigidos son próximos o inferiores al 15%, los errores de predicción aumentan de forma leve.
- Fijando el 30% como la probabilidad a considerar, los modelos experimentan el mayor grado de mejora, reduciendo sus tasas de fallo en más de seis puntos porcentuales.

Por ello, parece evidente como la estructura sectorial tienen un impacto muy significativo en el impago de la prima; además, lanza un mensaje de la importancia que pueden tener los aspectos económico-financieros de los tomadores en la explicación de este hecho.

Además, se ha dotado de una calificación a las predicciones de este nuevo conjunto de modelos, al igual que se hizo con los modelos iniciales, para analizar si las pólizas que tienen un determinado rating, dada una probabilidad de impago estimada por el modelo, finalmente termina incurriendo en impago. En este sentido, los resultados obtenidos son los que se recogen a continuación:

	Tasa de Impago Según Calificación											
	Modelo 13		Modelo 14		Modelo 15		Modelo 16		Modelo 17		Modelo 18	
Rating	Logit	Probit	Logit	Probit	Logit	Probit	Logit	Probit	Logit	Probit	Logit	Probit
A	3,3%	2,9%	3,0%	3,0%	3,1%	3,0%	3,1%	3,2%	2,9%	2,9%	2,9%	2,8%
B	4,9%	5,0%	5,7%	5,6%	5,8%	5,7%	5,9%	5,7%	5,7%	5,6%	5,8%	5,7%
C	10,2%	10,2%	10,3%	10,3%	10,2%	10,3%	10,2%	10,2%	10,2%	10,3%	10,2%	10,3%
D	18,0%	18,1%	17,6%	17,5%	17,6%	17,5%	17,6%	17,7%	17,7%	17,7%	17,6%	17,7%
E	31,3%	30,9%	31,8%	32,1%	31,7%	32,0%	31,6%	31,5%	31,8%	31,6%	31,7%	31,5%
F	26,6%	23,7%	41,6%	41,0%	42,0%	41,8%	41,8%	42,7%	40,7%	41,5%	41,3%	41,3%

Como se observa, y en línea con lo anterior, la utilización de un sistema de rating como el presentado nos dota de una herramienta capaz de predecir el número de pólizas que no van a pagar la prima, en los términos establecidos, una vez estas han sido clasificadas en función de su probabilidad de impago estimada. En este sentido, los resultados obtenidos son parecidos entre los modelos; pues tras la clasificación de las pólizas, en función de su rating asociado, se logra predecir con bastante precisión el número que finalmente incurrirá en impago dentro de cada clase.

Centrando el análisis en el modelo 18, se han encontrado tasas de impago muy similares tanto para la metodología logit como la probit. Aquellas pólizas con la mejor calificación (Rating A) no pagan la prima en el 2.8% - 2.9% de los casos, cuando se contemplaba un rango hasta el 3%; lo que es destacable al estimar con precisión hasta aquellas pólizas con probabilidades muy reducidas. Para el resto de las calificaciones, los impagos acontecidos se encuentran dentro de los intervalos deseados. De esta forma, la proporción de pólizas con rating B que no pagarían la prima se encuentra en cifras próximas al 6%, cuando se contemplaban valores comprendidos entre el 3% y 7.5%, del 10% para las pólizas con rating C, cuyo rango abarca desde el 7.5% hasta el 12.5%, inferiores al 18% para las tipo D y cuyo intervalo lo conformaban tasas en el intervalo 12.5% - 25%. En cuanto a las peores calificaciones, que contemplaban probabilidades de impago entre el 25% y el 40% en el caso del rating E o superiores al 40% para las F, se ha encontrado tasas del 32% y 41% respectivamente.

A la luz de los resultados obtenidos, estamos en disposición de afirmar que existen otras vías de estimación de probabilidades de impago. De hecho, la forma en que estas se obtendrían no precisan de solicitar ningún tipo de información adicional a los clientes, ni contratar los servicios de un proveedor externo que aporte estos datos a la entidad aseguradora, con los gastos implícitos que conlleva. Además, se confirma la relevancia de los indicadores macroeconómicos en los niveles de impago de las distintas provincias del territorio nacional, junto con el impacto que tienen cuestiones como la forma en que esta es pagada y su cuantía, el trimestre de contratación o su canal de emisión, entre otros. Dado que no resulta realista considerar que la totalidad de pólizas con una probabilidad estimada superior a un target fijado incurrirán en impago y todas las demás si pagarán, se ha encontrado un sistema de calificación capaz de definir su calidad crediticia, de forma que se contempla el impago para todas las probabilidades estimadas, si bien en función de la nota obtenida, su propensión al impago será diferente; este se ha contrastado con el comportamiento esperado en términos de impago para cada rating, coincidiendo con los valores deseados. El potencial de estos resultados es muy alto, pues permitiría conocer en el momento de cotización de la póliza cuál es su riesgo de impago asociado, prestándose a ser recalificada en periodos posteriores en base al comportamiento observado (por antigüedad). Del mismo modo, y como se comentará en próximas líneas, permitiría conocer cuáles son las palancas que lo potencian, para poder establecer mecanismos que reduzcan la exposición de la entidad.

4. Cómo reducir su impacto

Una vez conocidos los niveles de impago que se producen en la cartera, y algunas variables que pueden ayudar a predecir este fenómeno, sería interesante analizar distintas alternativas que permitan disminuir su repercusión, bien sea a través de factores asociados a las características de la póliza, contemplados en el modelo, como con la introducción de medidas contra el impago dentro del condicionado del contrato.

Prima de Riesgo

Al igual que sucede con los activos financieros, las compañías generalmente ofrecen distintas condiciones contractuales, en función de la probabilidad de que lo acordado entre las partes se lleve a término, en los plazos y forma pactados. De esta forma, las entidades tratan de controlar la calidad de sus clientes, para evitar situaciones en que la contraparte no pague el capital acordado, como pudiera ser la no devolución de un préstamo bancario, y reducir así su exposición. Tras lo ocurrido en el mercado de crédito hipotecario y de bonos del Estado, que terminó desembocando en la crisis bancaria y de deuda soberana, se comprobó cómo los tipos de interés exigidos por títulos a corto y medio plazo experimentaron un gran aumento, asociado esencialmente al riesgo de que a su vencimiento, las entidades y Estados emisores no pudieran hacer frente a su pago. Este hecho, por el que una mayor probabilidad de impago se traduce en unos tipos de interés más elevados, es lo que definiremos como spread, diferencial de crédito o prima de riesgo; que vendría a expresar cuanto más se tiene que compensar al prestatario por asumir un riesgo mayor ante contratos que en esencia tienen características similares.

Trasladando lo anterior al campo asegurador y, particularmente, al caso objeto de estudio, sería interesante considerar la posibilidad de introducir un spread de crédito asociado al riesgo de impago que presentan las pólizas. Para ello, haríamos uso de los modelos previamente estimados, que permitirían calcular esta probabilidad para cada póliza, tanto en el momento de su emisión como en los años que permanezca en cartera, suministrando el rating propuesto en líneas anteriores, y asignar un recargo sobre la prima en función de la calificación obtenida.

Este sistema, muy similar al empleado en el área de crédito del sector bancario, lograría compensar las pérdidas asociadas a los posibles impagos contemplados, con los recargos aplicados a las pólizas, según su riesgo.

Métodos de Pago

Además de la imputación de una prima de riesgo, existen vías ligadas a la contratación que pueden reducir las probabilidades de que una póliza no pague la prima. Para ello, profundizaremos en el análisis de coeficientes del modelo previamente estimado (modelo18). Cabe destacar que muchas de las variables que incluye no pueden ser alteradas o inducidas por la entidad pues si, por ejemplo, una póliza se emitió en la zona sur, no tiene lógica considerar un cambio de región de emisión; sino que deberíamos pensar en el spread de crédito como vía para penalizar su mayor probabilidad de impago. Sin embargo, la modalidad y el método de pago si son factores que la entidad puede controlar.

Por ello, a continuación se analiza el impacto que tendría, en términos de probabilidad de impago, que este sea fraccionado o único, o su posible domiciliación bancaria. A tal efecto, y dadas las características de la metodología de modelización propuesta, analizaremos el impacto en base a las diferencias que una observación presenta si alteramos cada una de las variables de manera independiente y si lo hacemos de forma conjunta, manteniendo el resto de los factores constantes.

Para ello, se han seleccionado un grupo reducido de pólizas, cuyas modalidades de pago eras fraccionadas, y que podían estar o no domiciliadas. De esta forma, se han realizado los análisis de sensibilidad que se detallan a continuación:

- Si tenía pago fraccionado y estaba domiciliada se ha calculado su nueva probabilidad de impago esperada en caso de que fuera pago anual.
- Si su forma de pago era fraccionada y además, no estaba domiciliada, se ha calculado su probabilidad de impago en caso de ser pago anual o de estar domiciliada o de que se dieran las dos condiciones al mismo tiempo.

Cabe destacar también las dos formas en que se presentaba la modalidad de pago: si la periodicidad era anual, semestral, trimestral o irregular, quedando contemplada por el modelo 16; o si se encontraba agrupada en función de si el pago era único o fraccionado, recogido por el modelo 18. Por tanto, en este análisis se emplearán ambas consideraciones para contemplar el impacto, en términos de probabilidades de impago.

Atendiendo al primero de los puntos descritos, encontramos como la modificación de la metodología de pago semestral a la anual, manteniendo el resto de los factores constantes, genera importantes reducciones en las probabilidades de impago:

	Modelo 16		Modelo 18	
	Logit	Probit	Logit	Probit
Inicial	15,07%	15,44%	15,05%	15,12%
Nueva	9,95%	10,09%	9,90%	9,88%
Reducción	5,13%	5,35%	5,15%	5,24%

Para otra de las observaciones, en esta misma línea, se ha analizado el impacto en su probabilidad de impago estimada, si pasase de la modalidad de pago trimestral a la anual, ceteris paribus:

	Modelo 16		Modelo 18	
	Logit	Probit	Logit	Probit
Inicial	13,19%	13,06%	12,87%	12,99%
Nueva	8,15%	8,03%	8,39%	8,32%
Reducción	5,05%	5,03%	4,48%	4,67%

Al igual que lo acontecido para el caso semestral, la probabilidad de impago de la póliza analizada sería muy inferior si tuviese como modalidad la anual, reduciéndose en una cuantía de cinco puntos porcentuales, respecto al pago trimestral original.

De esta forma, comprobamos como el fraccionamiento del pago es una variable muy significativa para pólizas domiciliadas y cómo la modalidad anual debería ser la opción propuesta, para aquellas pólizas cuya probabilidades de impago estimadas sean superiores al nivel deseado o asumible.

A continuación y en respuesta al segundo punto, se ha llevado a cabo el mismo análisis pero tomando en consideración pólizas no domiciliadas. Con ello se busca analizar si la domiciliación aumenta o disminuye la probabilidad de impago estimada y cuál sería el impacto de que el pago pasase a ser único y estuviese domiciliado.

	Modelo 16		Modelo 18	
	Logit	Probit	Logit	Probit
Inicial	7,14%	7,02%	6,80%	6,65%
Nueva	9,07%	8,97%	8,87%	8,75%
Reducción	-1,93%	-1,95%	-2,07%	-2,10%

Como se puede observar, el pago domiciliado tiene un efecto reducido pero negativo o, lo que es lo mismo, la probabilidad de impago aumentaría si el pago de la póliza se realizase a través de domiciliación bancaria. La principal causa a la que se le puede atribuir este hecho se encuentra en la relación personal que suele existir entre agente y tomador en el caso de pólizas no domiciliadas, frente al caso contrario, dónde con la devolución del recibo o el vaciado de la cuenta sería suficiente para incurrir en impago.

Pese a ello, y dada la continua expansión de los medios de pago electrónicos y la contratación sin intermediación física, si parece interesante analizar este impacto en conjunción con el pago anualizado. Como recoge la siguiente tabla de resultados, el efecto agregado supondría una reducción de la probabilidad de impago próxima a un punto y medio porcentual; si bien estimando únicamente los efectos parciales para la misma póliza, la modificación del pago fraccionado semestral por el anual implicaría una disminución de la probabilidad de tres puntos porcentuales, frente al incremento del dos y medio que supondría la domiciliación bancaria del mismo.

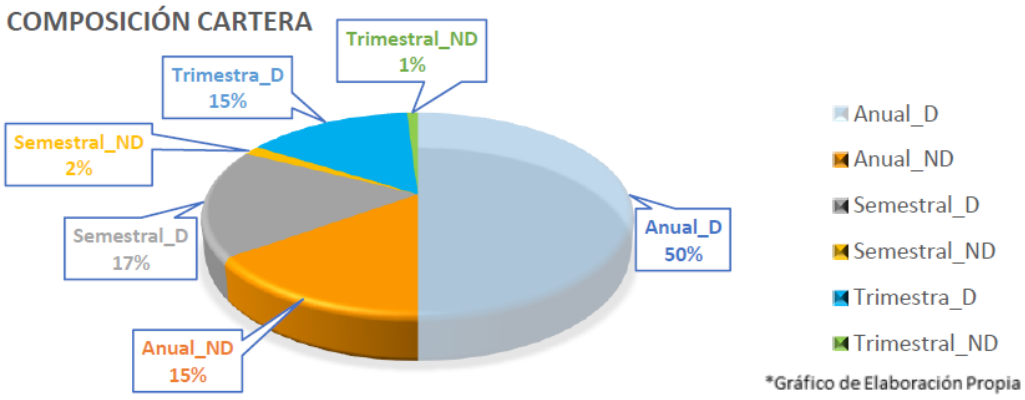
	Modelo 16		Modelo 18	
	Logit	Probit	Logit	Probit
Inicial	8,57%	8,60%	8,43%	8,24%
Nueva (Fraccionamiento)	5,51%	5,21%	5,40%	4,99%
Reducción por Fraccionamiento	3,06%	3,39%	3,03%	3,25%
Nueva (Domiciliación)	10,84%	10,85%	10,94%	10,69%
Reducción por Domiciliación	-2,27%	-2,26%	-2,50%	-2,45%
Nueva Total	7,03%	6,77%	7,08%	6,67%
Reducción Global	1,54%	1,83%	1,36%	1,56%

Realizado el mismo análisis para una póliza trimestral que no estaba domiciliada, se ha obtenido un efecto global del tres puntos porcentuales, con un efecto reductor próximo al cinco y medio con el cambio de modalidad de pago trimestral al anual, junto con un aumento cercano al tres y medio por la domiciliación del pago.

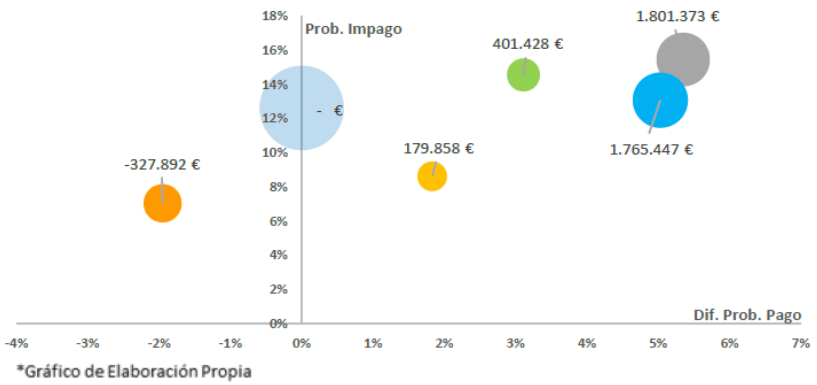
	Modelo 16		Modelo 18	
	Logit	Probit	Logit	Probit
Inicial	14,14%	14,54%	13,78%	14,43%
Nueva (Fraccionamiento)	8,77%	9,08%	9,02%	9,37%
Reducción por Fraccionamiento	5,38%	5,46%	4,76%	5,06%
Nueva (Domiciliación)	17,61%	17,76%	17,57%	17,99%
Reducción por Domiciliación	-3,46%	-3,22%	-3,79%	-3,57%
Nueva Total	11,09%	11,43%	11,67%	12,05%
Reducción Global	3,06%	3,12%	2,11%	2,38%

La potencia de este análisis de sensibilidad se encuentra en la forma en que se estructuran las carteras de seguros. A modo de ejemplo, a continuación se muestra una posible composición de cartera, junto con los efectos que tendría la aplicación de las medidas analizadas. Cabe destacar que, para un mayor acercamiento a la realidad aseguradora, el volumen de pólizas con pago anual será el predominante, junto con la domiciliación de este. Suponiendo que las observaciones previamente analizadas fueran la referencia de las distintas posibles combinaciones de fraccionamiento y forma de pago, se puede obtener cuanto es el valor económico del modelo, en términos de cantidad de prima neta que se ha conseguido cobrar con su introducción.

De tal forma, la cartera que se ha simulado estaría compuesta por pólizas de las siguientes características:



Dado que el análisis está orientado a contratos que no están domiciliados (ND) o cuyo pago está fraccionado (Semestral, Trimestral); el siguiente gráfico muestra la conjunción entre la tasa de impagos inicial, la reducción que esta experimenta al modificar las variables de pago anteriores, el volumen de prima neta que genera cada tipología y, por último, la diferencia en términos de prima neta que se cobraría con la aplicación del modelo.



Como se puede observar, y en línea con los resultados previamente comentados para las pólizas referencia[7], la probabilidad de impago para pólizas anuales no domiciliadas aumenta en dos puntos porcentuales, lo que supondría una pérdida teórica de 300.000€ con su domiciliación; sin embargo, los beneficios que ofrecen el resto de pólizas, excepto las anuales domiciliada que componen el 50% de la cartera y se mantienen constantes, serían muy elevados, destacando aquellas pólizas cuyos pagos se encuentran domiciliados pero son de carácter fraccionado ya que, además, de representar en un 32% las pólizas que componen la cartera, la dimensión de las burbujas que las representan son las más importantes tras la anual domiciliada. Con todo ello, el volumen total de prima neta cobrada, para las cien mil pólizas que componían esta cartera simulada, tras la modificación de la forma de pago sería de 3.820.213€.

Por ello, resulta evidente como la metodología de pago es un aspecto relevante en los niveles de prima que finalmente cobran las entidades. Dado que la realidad económica muestra una tendencia hacia la domiciliación bancaria, lo que supone contemplar un aumento de los impagos que se producirán en el futuro; las entidades deben centrarse en fomentar el pago anualizado de la prima, estableciendo mejores condiciones de contratación a los clientes que tomen la forma anual respecto a la fraccionada. La vía más sencilla de implementarlo sería informando al cliente, en caso de que eligiese el pago fraccionado, de las ventajas que podría obtener con su anualización, como podría ser un descuento sobre la prima neta a cobrar, intentando inducirle a esta modalidad. Pues es mejor para la compañía realizar un leve descuento sobre la prima a cobrar y garantizando su cobro que exponerse a que el cliente finalmente no la pague.

Seguro de Crédito

Junto a las vías que se han explorado para reducir o compensar las pérdidas potenciales que puede experimentar la compañía, la primera de ellas orientada al cobro de una prima de riesgo y la segunda ligada a las condiciones de contratación de la póliza, el mercado financiero pone a disposición de las entidades nuevas formas de transferencia del riesgo de crédito asociado a operaciones como pueden ser la adquisición de un préstamo hipotecario, de títulos de deuda o de venta y prestación de bienes y servicios.

[7] Elegidas de forma aleatoria para evitar revelar información sensible de la compañía

Entre este tipo de estrategias de coberturas caben destacarse dos:

- Seguros de Crédito: cuya finalidad es garantizar que una persona o entidad cobrará las cantidades acordadas en los términos establecidos en el contrato, a cambio del pago de una prima de seguro. En el caso de estudio se basaría en contratar un seguro con otra entidad que cubriera el riesgo de crédito de las pólizas en cartera para que, en caso de impago, esta entidad cubriera la prima adeudada, haciéndose posteriormente cargo su cobro a través de procesos administrativos o requerimientos judiciales.
- Credit Default Swaps (CDS): también conocidas como "permutas de incumplimiento crediticio", se trata de un derivado financiero negociado en el mercado Over The Counter (OTC). Su funcionamiento se basa en un comprador del instrumento que paga periódicamente una cierta cantidad de dinero (spread) a cambio de que el vendedor del título, en caso de que el activo subyacente impague, haga frente a las cantidades adeudadas.

Dado que se ha establecido un rating capaz de clasificar las pólizas en función de su probabilidad de impago, los CDS podrían ser una buena vía para asegurar el cobro de pólizas que tengan una determinada calificación. En el caso de los seguros de crédito, podría ser interesante su utilización para pólizas con primas elevadas. El principal problema que presentan estas vías de cobertura y mitigación del riesgo es el coste que suponen a la entidad, tanto en términos administrativos como por la adquisición del seguro de crédito o la permuta de incumplimiento crediticio, si bien podría ser financiado con el cobro de una prima de riesgo a los asegurados con mayores probabilidades de impago.

5. Áreas de Investigación Futura

Técnicas de Machine Learning

A pesar de que los resultados obtenidos permiten una correcta estimación de la probabilidad de impago de la prima de una póliza, junto con la evaluación de los factores que suelen tener un mayor impacto en este hecho, son numerosos los métodos de modelización existentes para la realización de este tipo de análisis. Gracias al desarrollo de los equipos informáticos y de softwares específicos para el tratamiento de grandes volúmenes de datos, se ha abierto un nuevo campo de investigación enfocada en extraer la máxima información posible de los enormes repositorios que poseen las compañías sobre sus clientes.

Es en este punto dónde surge la figura del "data scientist" como la persona capaz de desarrollar algoritmos matemáticos combinados con técnicas computacionales que, aplicados sobre una determinada base de datos, generalmente grande y en la que los estos pueden o no estar estructurados, son capaces de extraer que factores son significativos en el comportamiento de la variable de análisis, desarrollando modelos con altas capacidades predictivas.

Dentro de las técnicas que conformarían parte de la denominada "inteligencia artificial", aparecen los métodos de "machine learning". Estos se pueden definir como programas informáticos que a través de procesos matemáticos logran predecir el comportamiento futuro utilizando la información inicialmente disponible y toda aquella que se vaya generando en momentos futuros, teniendo así un carácter de aprendizaje dinámico.

En la práctica actuarial, los métodos de machine learning están en continuo desarrollo y su aplicación se ha extendido a la mayoría de las entidades del sector, dada la potencia de los algoritmos que estas emplean y su gran capacidad de modelización de comportamientos futuros de individuos y variables. Aunque son numerosos los métodos de machine learning existentes, a continuación se presentan las nociones teóricas de aquellos que pudieran ser interesantes tener en cuenta en el futuro:

Árboles de Clasificación

Dado el carácter binario de la variable de interés y el de la mayor parte de los factores incluidos en el análisis, se ha encontrado en los árboles de clasificación otra forma de explicar la probabilidad de impago de las pólizas. Esta metodología, empleada tanto en el área de economía de la empresa, con la teoría de juegos, como en las técnicas de inteligencia artificial, se basa en la realización de diagramas con construcciones lógicas que, fijando una variable objetivo inicial, agrupa las observaciones.

Dentro del campo de la inteligencia artificial y el machine learning, este proceso de clasificación se vuelve mucho más complejo, pues el orden en que se dividen los grupos se basa en algoritmos matemáticos que detectan las características comunes que guardan las pólizas de la muestra inicial, dividiéndolas en grupos y, atendiendo a las diferencias que las integrantes de estos nuevos grupos generados, repetir el proceso en función de las características de las observaciones que lo contiene; por lo que la variable de división en cada subgrupo puede diferir entre ellos, pese a estar en el mismo nivel. Además, se pueden establecer numerosas limitaciones, como pudiera ser la clasificación en origen en torno a una variable particular. Uno de los principales criterios de agrupación es el basado en el concepto de entropía, seleccionando como variable de clasificación aquella que logre reducirla entre las observaciones que componen cada grupo; ya que lograría aglutinar en grupos cada vez más homogéneos a una muestra inicial heterogénea.

Dada la existencia de softwares que permiten la realización de este tipo de análisis de forma relativamente sencilla, a continuación se muestra un ejemplo real generado a través del paquete estadístico R, en el que se define una variable "$Y2$" como variable objetivo y que será la inversa de "$Y1$", es decir, si la póliza pagó la prima toma el valor uno y si incurrió en impago el valor cero.

Es importante conocer el criterio establecido en el algoritmo para la clasificación de observaciones, en este caso basado en la "impuridad de Gini"[8], medida de frecuencia sobre las ocasiones en que una observación elegida aleatoriamente ha sido etiquetada de forma aleatoria y errónea en un subgrupo, atendiendo a las características de este. Su cálculo, para una serie de elementos cuyas variables toman los valores $\{1,2,\dots,m\}$ y

[8] https://es.wikipedia.org/wiki/Aprendizaje_basado_en_%C3%A1rboles_de_decisi%C3%B3n

dónde "t_i" representa la tasa o fracción de elementos que se han etiquetado con valor "i" en el conjunto, se obtiene a través de la siguiente expresión:

$$I_G(t) = \sum_{i=1}^{m} t_i(1-t_i) = \sum_{i=1}^{m}(t_i - t_i^2) = \sum_{i=1}^{m} t_i - \sum_{i=1}^{m} t_i^2 = 1 - \sum_{i=1}^{m} t_i^2$$

Además, también es importante evaluar el grado de profundidad que se le quiere dar al árbol de clasificación, definida en el código a través del parámetro de complejidad "cp"; vendría a expresar el grado de explicación que se genera introduciendo más pasos al árbol, de forma que la máxima representación se encontraría para un valor "$cp = 0$".

A continuación se muestra el código empleado para la obtención del árbol de clasificación que se analizará en las próximas líneas, cuyo objetivo no es otro que mostrar otras vías alternativas a los modelos lineales generalizados anteriormente presentados, y la intuición que hay tras ellos. Además, dada la gran dimensión que su representación gráfica puede llegar a alcanzar, se ha empleado una muestra reducida compuesta por siete mil quinientas observaciones de una de las muestras aleatorias previamente generadas.

Código:

Se instalan y ejecutan los paquetes necesarios para la realización de árboles de clasificación y su posterior representación gráfica.

```
Install.packages("rpart","rpart.plot")
library(rpart)
library(rpart.plot)
```

Se genera el árbol, estableciendo la variable objetivo (Y2) y el conjunto de variables explicativas deseadas. Dado que queremos un árbol de clasificación establecemos el método "class" (existe la posibilidad de hacer un árbol de regresión para variables continuas). Introducimos los datos, 7500 observaciones en este caso, y fijamos un 0.27% como parámetro de complejidad.

```
CART_1 < -rpart(Y2~X4 + X9 + X12 + X20 + X22 + X23 + X27 + X28 + X29 + X33 + X96 +
X90 + X92 + X93, method = class, data = BBDDML75K[1:7500,], cp = 0.0027)
Windows()
Rpart.plot(CART_1, extra = 104, type = 1, box.pallete = "RdYlGn", branch.lty = 1, branch.col
= ("firebrick","palegreen2")[CART_1$frame$yval], shadow.col = "gray", nn
= TRUE, fallen.leaves = TRUE, faclen = 0, varlen = 0)
```

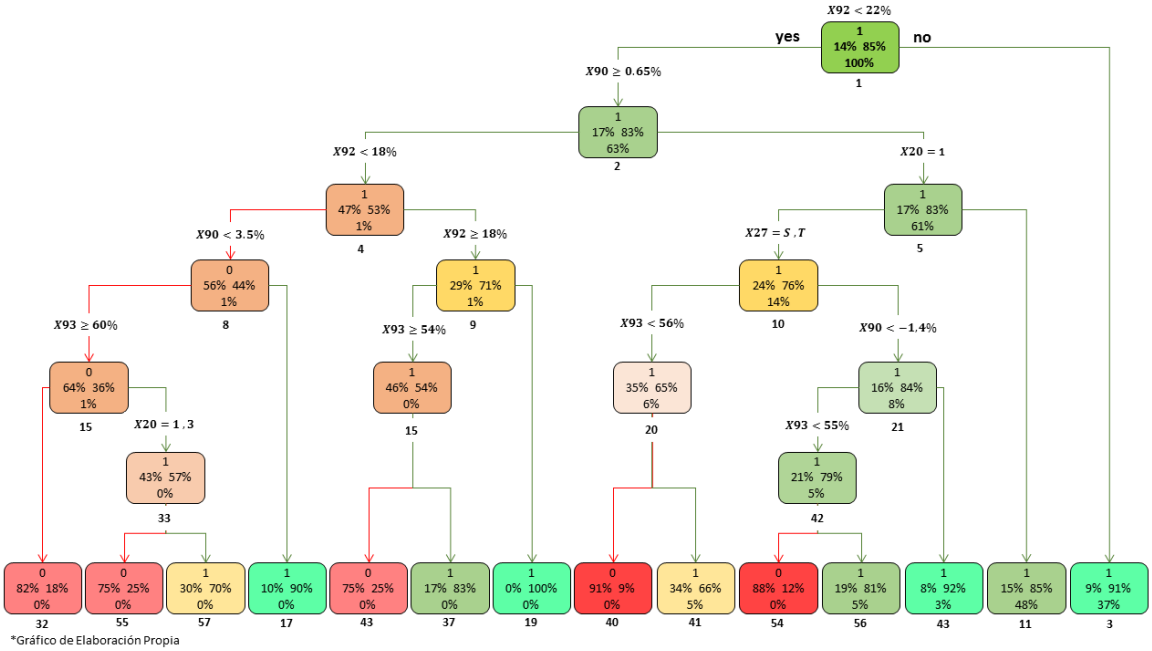


*Gráfico de Elaboración Propia

El gráfico presentado muestra el proceso de clasificación de pólizas siguiendo el criterio previamente descrito. De esta forma, en cada paso, muestra la variable clasificadora considerada y el valor que esta debe tomar para pertenecer a uno u otro grupo. Además, dentro de cada cuadro, recoge el valor que está considerando para la variable objetivo Y2, la proporción de pólizas dentro del grupo que toman ese valor y el porcentaje de observaciones que se encuentran dentro del mismo respecto al total de pólizas que inicialmente componían la muestra. Los colores empleados expresan el grado de pureza u homogeneidad que existe dentro de las pólizas que cumplen los criterios de esa rama, en términos de pagos o impagos que en estos se producen; de esta forma, el verde claro expresaría que casi la totalidad de pólizas que se encuentran en este pagan la prima, frente al rojo oscuro en que la mayoría incurrirían en impago.

Entrando en el análisis del diagrama, se encuentra un 85% de tasa de pago en la muestra analizada y cómo es el peso del sector secundario en el PIB de la región (X92) el primer clasificador, de forma que aquellas provincias que presenten valores mayores al 22% irían por la rama derecha del árbol, lo que implicaría unas tasas de pago del 91%, representando además el 37% del total de pólizas de la muestra analizada. En el caso de regiones con un peso inferior al mencionado, sería del 83% la tasa de pago, siendo la variación de la tasa de paro de la región la siguiente variable clasificatoria. Este proceso continua hasta llegar al final del árbol, dado el parámetro de complejidad establecido.

Entre las conclusiones que podemos extraer del ejemplo propuesto, es cómo las variables macroeconómicas son un factor muy relevante en la clasificación de pólizas y sus probabilidades de pago/impago, pues el 91% de aquellas situadas en provincias con un peso del sector secundario igual o superior al 22% pagan la prima, al igual que sucede en el 85% de los casos en que, pese a no darse la condición anterior, si eran pólizas con antigüedad superior al año. En caso de que llevara un año o menos en cartera, será la forma de pago la siguiente variable de clasificación; dónde si es distinta al semestral o trimestral encontraremos que, si la variación del paro es menor al -1.4% (estando entonces comprendido entre el -1,4% y el 0,65%), en el 92% de los casos la prima se paga; en caso de no darse esta condición, será importante conocer el peso del sector terciario, pues si es inferior al 55%, la póliza incurrirá en impago en el 88% de los casos.

Mención especial merece el caso en que la probabilidad de pago es máxima, formando lo que se conoce como clase pura, en la que el 100% de las pólizas pagarían la prima (situadas en la clase 19); o el caso contrario, en el que encontramos la cota máxima del impago, dentro del grado de depuración o complejidad contemplado, dónde para aquellas pólizas con un peso del sector secundario menor al 22%, con variaciones del paro inferiores al 0.65%, cuya antigüedad es igual o menor a un año, en las que el pago se realiza de forma semestral o trimestral y cuyo sector terciario tiene un peso menor al 56%, poseen unas tasas de impago superiores al 90%.

Por tanto, este sistema permite visualizar la distribución del impago para pólizas que cumplen una determinada serie de características, tanto propias como de la situación económica en que se encuentran enmarcadas. Destacan aquellas que se encuentran en los nodos 40 y 54 por tener las tasas de impago más elevadas, frente al grupo 19 en el que encontramos una clase pura dónde la totalidad de ellas pagan la prima, u otros como son los grupos 17, 43 y 3, con tasas de pago superiores al 90% de los casos.

De forma adicional, con la introducción de las sentencias mostradas a continuación, se puede observar también el tipo de relación que guardan algunas de las variables recogidas por el árbol tanto de manera individual como conjunta, respecto a la probabilidad de pago de la prima. Esta relación será de carácter bidimensional en el primero de los casos y tridimensional para el segundo, pues se estará introduciendo una combinación de dos variables (X_1, X_2) para obtener el valor de la objetivo (Y2).

Código:

Se instala y ejecuta el paquete necesario para la realización de estas gráficas.

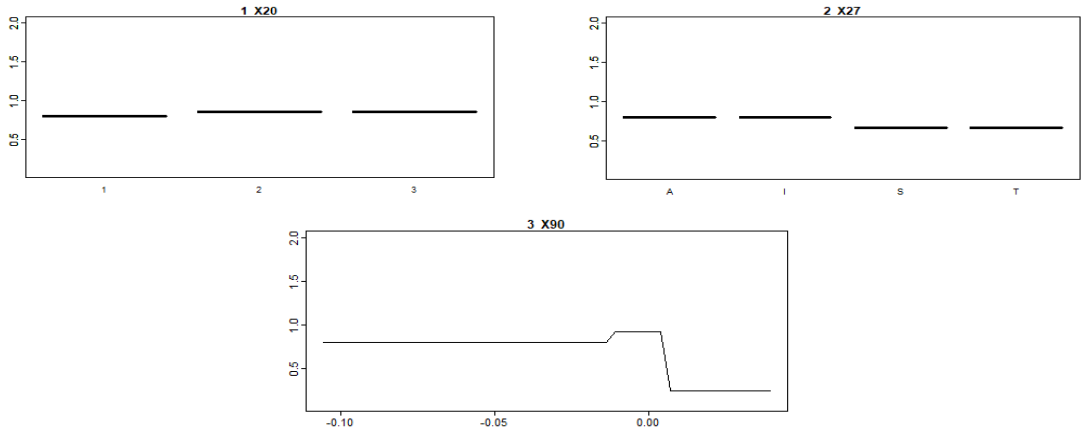
```
Install.packages("plotmo")
library(plotmo)
```

Generamos el modelo de respuesta para el rango de valores predictores, haciendo uso del árbol de clasificación previamente calculado, definiendo que la variable objetivo mide una probabilidad, la de pago (Y2) en este caso.

```
Windows()
plotmo (CART_1, type = "prob", col.persp = "cornflowerblue")
```

La potencia de esta herramienta se encuentra en la capacidad para intuir posibles comportamientos de la variable objetivo en función, por ejemplo, del valor que pueda tomar una variable construida como factor. A tal efecto, los gráficos mostrados a continuación son únicamente aquellos que se ha considerado pueda ser interesante analizar, si bien en el anexo 3 se encuentran la totalidad de ellos.

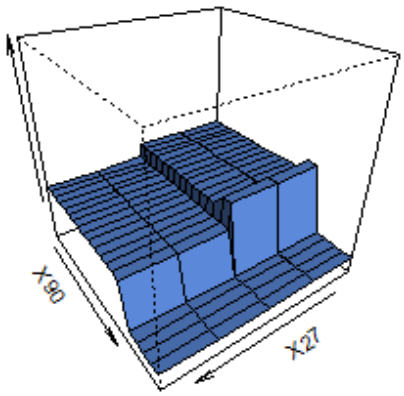
Atendiendo a las representaciones univariantes, se encuentra como las pólizas con una antigüedad (X20) mayor en cartera poseen tasas de pago superiores, contrario a lo que sucede en el caso de que su forma de pago (X27) sea semestral o trimestral. En relación con las variables macroeconómicas introducidas aparece, en relación con la tasa de variación del paro (X90), dos efectos: el primero, como era previsible, muestra como los niveles de pago disminuyen con aumentos del paro (tasas positivas), sin embargo, surge también un detalle muy relevante por el que pequeñas reducciones del paro tienen un efecto, en términos de pago, mayores que grandes disminuciones de este. La intuición detrás del hecho puede encontrarse en la gran volatilidad de estas regiones y el impacto del ciclo económico en el que se encuentre la economía nacional.



Pasando al análisis de los gráficos multivariantes, cabe destacar la forma en que estos quedan dispuestos, dónde la base de la forma cúbica en que se dispone contendrá los valores que toman las variables (X_i, X_j), y el sentido en que estas se disponen, siendo la altura el valor de Y asociado. De esta forma, se encuentran diferencias entre la interacción de variables para la totalidad de ellas, si bien las más importantes son las que se producen entre la variación de la tasa de paro (X90), la forma de pago (X27) y las demás variables consideradas; cabe recordar que el árbol había realizado la división en base al pago semestral y trimestral, frente a las otras modalidades de pago.

Pasando al análisis multivariante, los gráficos quedan dispuestos de forma cúbica, dónde la base contiene los valores que toman las variables (X_i, X_j), y el sentido en que estas se disponen, siendo la altura el valor de Y asociado, es decir, su probabilidad de pago. De esta forma, se encuentran diferencias entre la interacción de variables para la totalidad de ellas, si bien es la forma de pago (X27) aquella en que nos detendremos, en línea con el sentido del estudio. Cabe recordar que el árbol había realizado la división en base al pago semestral y trimestral, frente a las otras modalidades de pago.

Si antes se mostró cómo las formas de pago semestral y trimestral ofrecen unas menores probabilidades de que se pague la prima frente a la anual o la irregular, es el siguiente gráfico en el que se observa cómo, en su interacción con la tasa de variación del paro (X90), las diferencias entre clases son mayores, pues las de estas modalidades, en el tramo negativo, ofrece mejores resultados. Sin embargo, son las comprendidas en el intervalo $(-1.4\% \, , 0.65\%)$ las que ofrecerían mayores niveles de pago.



Por todo ello, se ha encontrado en los árboles de clasificación un buen método de agrupación de pólizas atendiendo a sus características, logrando obtener, en algunos casos, clases casi perfectas en términos de tasas de pago e impago dentro de las pólizas que los componen. Además, ha servido para reafirmar la hipótesis inicial, por la que se puede predecir el impago con el empleo de las características del contrato a emitir o que ya formaban parte de la cartera; con un importante peso de la situación macroeconómica y la organización sectorial de la provincia en que se sitúa la póliza.

Árbol de Modelos Lineales Generalizados

Otra de las metodologías que pudiera ser interesante analizar es la aplicación de los modelos lineales generalizados en conjunción con un árbol de clasificación, para comprobar si las predicciones mejoran, respecto a los modelos iniciales.

El procedimiento para la realización de este sistema combinado sería sencillo, una vez comprendidos los árboles de clasificación. Pues se basa en la realización de un algoritmo capaz de construir un árbol como el anterior, que organice las variables en función de su capacidad de reducir la entropía de la muestra inicial, sobre el que se aplicaría, para cada una de sus ramas, un modelo lineal generalizado. Logrando así explicar el comportamiento hacia el impago de las observaciones que conforman cada uno de los subgrupos en que se han clasificado las observaciones de la muestra inicial total.

De forma orientativa, para comprobar el potencial de esta forma de proceder, se ha llevado a cabo la estimación de los modelos probit y logit para cada una de las ramas, en una versión reducida del árbol anterior, que se muestra a continuación, pero estableciendo como variable objetivo el impago, en línea con los modelos iniciales:

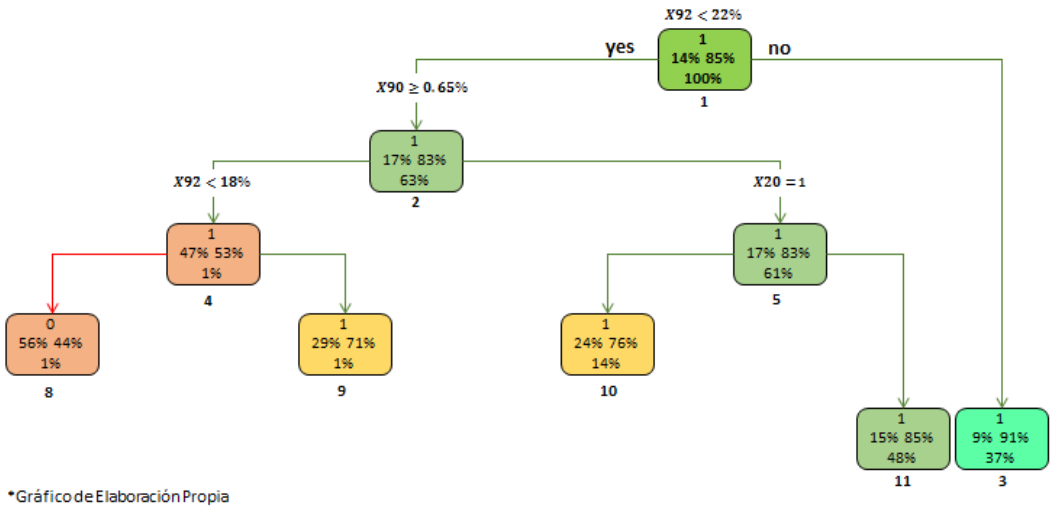


*Gráfico de Elaboración Propia

Por tanto, se ha estimado un GLMs para cada una de las siguientes ramas (adjuntos en el Anexo 4.1):

- Rama 1: Peso del sector secundario mayor o igual al 22%
- Rama 2: Peso del sector secundario menor que 22%, con tasa de variación del paro menor al 0.65% y que lleva más de un año en cartera
- Rama 3: Peso del sector secundario menor que 22%, con tasa de variación del paro menor que 0.65% y que lleva un año o menos en cartera, es decir, que es de nueva producción.
- Rama 4: Peso del sector secundario comprendido en el intervalo $(18\%\,,\ 22\%)$, con tasa de variación del paro mayor o igual que 0.65%.
- Rama 5: Peso del sector secundario menor que 18%, con tasa de variación del paro mayor o igual que 0.65%.

Dadas las características de cada póliza, el árbol otorgará una predicción de impago asociada a su correspondiente rama. Para la evaluación de la calidad de las predicciones que genera este procedimiento, se debe tener en cuenta la totalidad de predicciones, aciertos y fallos en todas las ramas. De esta forma, para la obtención de la tasa de fallos en la predicción del impago, se ha calculado el número de fallos en cada rama (Anexo 4.2) respecto del total de predicciones que estas contenían. Los resultados a tal efecto son los siguientes:

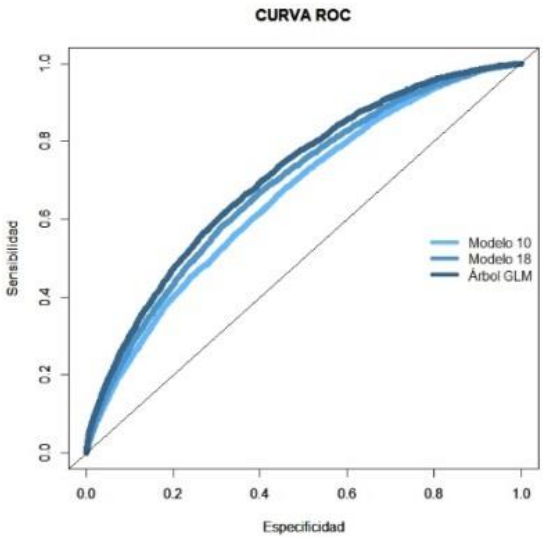
	GLM (Modelo 16)		Árbol GLM		Diferencias	
	Logit	Probit	Logit	Probit	Logit	Probit
40%	95,5%	96,3%	90,5%	91,3%	-5,3%	-5,2%
30%	83,7%	84,6%	79,4%	80,4%	-4,7%	-4,8%
20%	55,8%	55,2%	52,9%	52,4%	-3,4%	-3,0%
15%	34,4%	33,3%	33,4%	32,6%	-1,2%	-1,1%
10%	14,3%	13,6%	14,2%	13,7%	0,0%	0,2%
5%	2,1%	2,3%	3,1%	3,1%	1,0%	0,7%

Son evidentes las mejoras que supone la utilización de este tipo de modelización, frente al mejor de los modelos previamente estimados, pues se reducen en cifras próximas al cinco por ciento los fallos de predicción, considerando porcentajes de impago a partir del treinta y cuarenta por ciento. Todo ello, teniendo en cuenta que no se ha realizado para el mayor grado de depuración en que se encontraba representado, por lo que cabe pensar que una mayor profundidad de este con la fijación, incluso, de un parámetro de complejidad más pequeño al empleado puede dotarnos de predicciones de impago

mucho mejores que las aportadas por un único modelo lineal generalizado para la totalidad de la cartera. La matriz de confusión desarrollada para el árbol previamente descrito presenta notables mejoras, como se muestra a continuación:

Modelo	Probit	**Valor Predicho**		$\boldsymbol{VP} = 864$
Probabilidad	30%	Pago (0)	Impago (1)	$\boldsymbol{VN} = 23291$
Valor Real	Pago (0)	23291	1309	$\boldsymbol{FP} = 1309$
	Impago (1)	3557	864	$\boldsymbol{FN} = 3557$

En comparación al modelo con variables macroeconómicas, la capacidad de predicción de los casos de impago (VPR) aumenta en cinco puntos porcentuales, hasta el 19.54% con una leve disminución de capacidad de predicción asociada al pago (SPC), que se sitúa en el 94.67%, con una tasa de acierto (ACC) del 83.23%; lo que implica que el porcentaje de falsos positivos (FPR) se sitúe en el 5.32%. Además, los ratios predictivos positivos (PPV) y negativos (NPV) aumentan al 39.76% y 86.75% respectivamente, con un ratio de falsos descubrimientos (FDR) del 60.23%. Estas leves mejoras las observamos en la representación gráfica de las curvas ROC para los modelos 10, 18 y el árbol GLM.



Además, y en línea con la metodología empleada para la evaluación de los modelos, se ha calculado el nuevo rating, resultante de agregar la cantidad de pólizas que recibiría cada una de las calificaciones contempladas y cuántas de ellas realmente impagaron. Como se puede observar, las tasas de impago observadas en cada calificación se encuentran dentro de los criterios establecidos, comportándose de forma muy similar al mejor de los modelos previos:

	Modelo 18		Árbol GLM		Probabilidad de Impago
Rating	Logit	Probit	Logit	Probit	
A	3%	3%	3%	3%	$P(Y_i = 1) \in [0\,,0.03]$
B	6%	6%	5%	5%	$P(Y_i = 1) \in (0.03\,,0.075]$
C	10%	10%	10%	10%	$P(Y_i = 1) \in (0.075\,,0.125]$
D	18%	18%	18%	18%	$P(Y_i = 1) \in (0.125\,,0.25]$
E	32%	32%	31%	31%	$P(Y_i = 1) \in (0.25\,,0.40]$
F	41%	41%	49%	49%	$P(Y_i = 1) \in (0.40\,,1]$

Boostrapping[9]

Como se mencionaba en líneas anteriores, la cartera inicial estaba compuesta por más de cien mil observaciones, que fueron divididas de forma aleatoria en una muestra de setenta y cinco mil para la estimación de los distintos modelos presentados, utilizando las observaciones restantes para la realización de predicciones que aportaran evidencia empírica de la validez de los modelos.

La idea del boostrapping se centra en la forma en que es dividida la muestra ya que, si bien parece correcto la separación aleatoria de observaciones para los puntos de modelización y predicción y más, conociendo el elevado tamaño de la muestra inicial, no se debe omitir que los resultados obtenidos corresponden con la división realizada y que, si se repitiese el proceso, estos no serían exactamente los mismos, aunque se esperaría un comportamiento parecido.

Con esta nueva técnica basada en el remuestreo, se llevaría a cabo tantas veces como se considerase oportuno el proceso de división aleatoria de la muestra entre las observaciones destinadas para la estimación de modelos y las dedicadas a su validación, realizando todo el proceso que hemos desarrollado para el análisis de coeficientes y la

[9] http://www.sld.cu/galerias/pdf/sitios/revsalud/tesis_de_resampling.pdf

extracción de conclusiones sobre la capacidad predictiva de estos. Intuitivamente, el proceso a realizar sería el siguiente:

1. División aleatoria de la cartera en dos muestras (modelización y validación).
2. Estimación de modelos probit y logit con la muestra de modelización.
3. Realización de predicciones con los modelos estimados y la muestra de validación, obteniendo las probabilidades de impago.
4. Almacenar resultados.
5. Repetir el proceso desde 1 hasta 4 tantas veces como se desee, teniendo en cuenta los requerimientos operativos.
6. Con los resultados de los coeficientes estimados, generar una función de distribución de probabilidad asociada a cada uno de ellos.
7. Con el nuevo modelo, realizar predicciones de impago enfrentándolas con los valores observados.

De esta forma, se evitaría que los resultados obtenidos estuviesen influenciados por la muestra aleatoria empleada para su estimación, mejorando así la capacidad predictiva de los modelos y su validez. Además, permitiría valorar el error derivado de la realización de una única división aleatoria y establecer intervalos de confianza para los parámetros estimados.

Consideración de Variables Financieras

La forma en que el riesgo de default ha sido modelizado se ha centrado, principalmente, en aquellas características que presentaban las pólizas y no en la calidad crediticia de aquellos que deben hacer frente al pago de la prima, es decir, de sus tomadores.

Por ello, sería interesante incluir en el análisis aspectos financieros de las entidades aseguradas con el objetivo de conocer el impacto de estos sobre la probabilidad de impago. Dado que son numerosas las masas patrimoniales que organizan los activos y pasivos de las entidades e, individualmente, pueden no aportar una gran información, sería interesante incluir las relaciones existentes entre activos y pasivos mediante ratios financieros. Aunque habría que analizar cuáles de ellos aportan una mayor información y realizar análisis de multicolinealidad para evitar la introducción de variables altamente correlacionadas que puedan explicar efectos similares, los principales indicadores a contemplar serían:

- Apalancamiento Financiero: mide el nivel de endeudamiento con entidades financieras que la empresa posee en pasivos a corto y largo plazo respecto de su volumen de activos. Generalmente, este ratio está en función del plan estratégico de la compañía, pues si el coste del capital de pedir prestado a entidades de crédito es menor a la rentabilidad que le generan los activos asociados, permite a la entidad crecer sin reducir el rendimiento que obtiene el accionista.

$$Apalancamiento\ Financiero = \frac{Activo\ Corriente + Activo\ No\ Corriente}{Pasivo\ Con\ Entidades\ Financieras} \cdot 100$$

- Fondo de Maniobra: representa el volumen de activos a corto plazo que poseería la empresa una vez liquidados las obligaciones a corto plazo que tenía contraídas con sus acreedores. Sirve para comprobar el margen que posee la entidad en activos a corto plazo para poder hacer frente a sus obligaciones en caso de que los derechos destinados a su cobertura disminuyan su valor.

$$Fondo\ de\ Maniobra = Activo\ Corriente - Pasivo\ Corriente$$

- Prueba ácida: analiza la capacidad de la empresa para hacer frente a sus pasivos a corto plazo haciendo únicamente uso de los activos más líquidos (tesorería y bancos), sin tener que liquidar los activos que conforman su inventario de existencias, mercaderías… Es quizás, el ratio más empleado para evaluar el nivel de liquidez de una entidad.

$$Prueba\ Ácida = \frac{Activo\ Corriente - Existencias}{Pasivo\ Corriente} \cdot 100$$

- Ratio de Solvencia: estudia el grado en que el total de activos que componen la empresa pueden hacer frente a la totalidad de deudas contraídas con sus acreedores. De esta forma, no tiene en cuenta únicamente la situación a corto plazo como la prueba ácida, sino el global de derechos y obligaciones de la entidad. Como mínimo, las empresas deben mantener un ratio de solvencia superior al 100% para garantizar su viabilidad, aunque cuanto mayor sea el valor de este mejor será la situación de la compañía, reduciendo también los costes del capital asociados al riesgo de impago. En el caso de entidades aseguradoras, resulta frecuente encontrar ratios de solvencia superiores al 200%.

$$Ratio\ Solvencia = \frac{Activo\ Corriente + Activo\ No\ Corriente}{Pasivo\ Corriente + Pasivo\ No\ Corriente} \cdot 100 \geq 100\%$$

Una vía en que estos factores podrían ser introducidos en los modelos previamente estimados, con una base de datos que incluyera la información financiera y datos sobre impagos en el plazo acordado a entidades de crédito (no de la prima de seguro), tanto de las entidades aseguradas, como de otras con características similares, sería generando un modelo lineal generalizado cuya variable dependiente fuera la probabilidad de impago de la prima del seguro, y en el que se incluiría como regresores las variables independientes que hemos comprobado que eran estadísticamente significativas, junto con una variable que recoja la probabilidad de impago a entidades de crédito, estando esta última correlacionada con el término de error. De esta forma, se podría generar una regresión en dos etapas de forma similar al ejemplo lineal reflejado a continuación:

- Etapa 1: estimación de un modelo que relacione una variable "X_{k1}" que recoge la probabilidad de impago en su plazo establecido, con lo ratios que previamente se han mencionado y que servirán como instrumento. Por ejemplo, si definimos los instrumentos "Z_1" como el ratio de solvencia y "Z_2" como el apalancamiento financiero de la entidad, definiríamos la primera etapa de la forma:

$$P(X_{k1} = 1) = \pi_0 + \pi_1 \cdot RatioSolvencia + \pi_2 \cdot ApalancamientoFinanciero + v_i$$

- Etapa 2: una vez obtenida una probabilidad de impago a entidades de crédito se introduciría en el modelo de variables instrumentales, siendo "X_{k1}" un regresor más del modelo, pero que ya no estará correlacionado con el término de error al utilizar como instrumentos los ratios financieros. El ejemplo de la segunda etapa y continuando con la etapa 1, incluye como variable exógena el sector de actividad que, al ser un factor de tres valores, el primero lo engloba el intercepto y los sectores secundario y terciario los coeficientes β_1 y β_2

$$P(Y = 1) = \beta_0 + \beta_1 \cdot X4S2 + \beta_2 \cdot X4S3 + \cdots + \beta_{k1} \cdot X_{k1} + \cdots + u_i$$

La forma de entender esta metodología de estimación de coeficientes sería como un modelo lineal generalizado sintético, dónde uno o varios de los factores explicativos del modelo procede de otra serie de modelos lineales generalizados previamente estimados que han servido para parametrizarlos y capturar la correlación que estas guardaban con el término de error del modelo global, corrigiendo el sesgo por variable omitida con la utilización de instrumentos.

Behaviour Economics

En los últimos años, se está desarrollando dentro del campo de la economía lo que se define como economía conductual. Esta ciencia analiza el impacto que tiene la psicología y la sociología sobre el comportamiento de los individuos, con el objetivo de integrar en los modelos tradicionales factores que puedan ayudar a predecir el futuro con una mayor precisión. En este sentido, el Premio Nobel de Economía Richard Thaler ha encontrado un importante campo de estudio en "cómo las limitaciones en el raciocinio, las preferencias sociales y la falta de autocontrol afectan a las decisiones individuales y a las tendencias en el mercado"[10].

El motivo por el que se ha generado esta nueva línea de investigación se fundamenta en las limitaciones de la metodología clásica, basada en el comportamiento racional de los individuos en el momento de tomar una decisión. Tradicionalmente, para posibilitar la realización de inferencia y estimaciones de lo que sucedería en el futuro, se establecía un supuesto de racionalidad que implicaría un análisis de las distintas alternativas disponibles, la utilidad que le generan al sujeto… En realidad, a la hora de tomar una decisión, el individuo no es capaz de contemplar la totalidad de escenarios posibles, analizando cuál de ellos le reporta una mayor utilidad real y elegir así la mejor de las alternativas que se le planteaban; sino que en muchos casos responden a impulsos, hábitos o respuestas a los esquemas que la mente genera para su simplificación.

Cierto es que los nuevos sistemas de adquisición, almacenamiento y procesamiento de los datos han hecho posible este tipo de análisis para un gran número de industrias, destacando la de la alimentación, en el que plataformas como WalMart observaron como las ventas de dos productos muy diferenciados (pañales y cerveza), a determinadas horas del día, eran mayores si se situaban en el mismo sitio.

En el ámbito asegurador y en línea con la temática de estudio, sería interesante analizar la relación existente entre la frecuencia y la severidad siniestral respecto a la probabilidad de impago de la prima. A priori, deberían tener una mayor cantidad de siniestros, o siniestros más costosos aquellas cuyos tomadores tienen una probabilidad de impago mayor, capturando el comportamiento a través de las características de la póliza contratada: coberturas y garantías, actividad y dimensión de la empresa, primas y capitales asegurados, localización geográfica del riesgo y tomador…Este hecho se puede asociar a dos cuestiones:

[10] http://www.expansion.com/economia/2017/10/09/59db4970e2704e82778b45ee.html

- Una mayor siniestralidad suele implicar un aumento de la prima de renovación y el tomador es posible que termine por no pagar la prima y cambiar de compañía.

- Entidades con mejores sistemas de protección y prevención generalmente presentan menores niveles de siniestralidad o, ante el acaecimiento de un siniestro, la gravedad del mismo es notablemente inferior. Además, empresas que muestran un comportamiento enfocado a la no tenencia de siniestros, pese a estar asegurados, suelen tener menores probabilidades de impago.

Es importante destacar la heterogeneidad de una cartera de seguros y cómo sería interesante realizar este análisis conductual, previa organización en clúster de las pólizas, logrando extraer el comportamiento común entre contratos de similares características, pues grandes grupos empresariales, pese a tener siniestros podrían tener una probabilidad de impago baja, dada su dimensión y el elevado número de situaciones de riesgo que se pueden encontrar aseguradas.

6. Conclusiones

Como se comentó al inicio del presente documento, el estudio se centraba en la búsqueda de un sistema capaz de predecir los niveles de impago para el seguro de daños para empresas, en base a la información contenida en una cartera de referencia; logrando así evitar los costes asociados a la compra de una evaluación crediticia, acerca de los tomadores, generalmente realizada por agencias de calificación externas. Además, se fijó como objetivo analizar si indicadores macroeconómicos como la estructura sectorial o el paro tenían un efecto significativo en las probabilidades de impago. También, se consideró interesante realizar un análisis detallado sobre el impacto de los métodos de pago en la variable de interés, buscando distintas vías que permitiesen reducir el riesgo de crédito al que las compañías aseguradoras se encuentran expuesto, o formas en que este puede ser cubierto o compensado. Por último, se ha llevado a cabo una leve introducción a los algoritmos de machine learning, explicando el funcionamiento de algunos de ellos y ejemplificando levemente las mejoras que pueden aportar a los modelos clásicos.

Atendiendo a las conclusiones que se desprenden del informe, cabe destacar cómo la utilización de la información suministrada por el tomador en el momento de emitir la póliza, junto con otras cuestiones orientadas a los canales de cotización, mes en que se generó o las primas y capitales objetos de seguro, se estaría en disposición de construir un modelo capaz de predecir los niveles de impago de la prima que se van a producir en la cartera. Además, cuestiones como la forma de pago, el área geográfica o los años que lleva operando con la compañía tienen una gran repercusión sobre el impago.

De igual modo, se ha comprobado como el establecimiento de un sistema de definición del impago de carácter binario, en base a una probabilidad fija a partir del cual es considerado que este se produce, no es de gran utilidad para hacer predicciones sobre el comportamiento de la cartera; pues supondría que pólizas con una probabilidad de impago estimada inferior a esa cuantía siempre pagarían, distando enormemente de la realidad empresarial.

Por ello, y en línea con los sistemas de evaluación crediticia empleados en el sector bancario, se ha definido un rating que dota a cada póliza de una calificación $\{A, B, C, D, E, F\}$ en función de las probabilidades de impago estimadas por el modelo. Con esta nueva forma de evaluar las predicciones, se ha encontrado que el

comportamiento real de las pólizas que componían las distintas calificaciones se encuentra dentro de los niveles de impago esperados en cada una de ellas. Por tanto, se está en disposición de afirmar que este sistema es realmente útil para conocer la exposición de la entidad al riesgo de crédito, en función del volumen de contratos que se enmarquen en cada rating. Su potencial es aún mayor si se tienen en cuenta las distintas vías existentes para compensar el riesgo asumido, como podría ser la introducción de un spread de crédito sobre la prima o acudir al mercado financiero para la contratación de un seguro de crédito o un CDS que permita transferir el riesgo a un tercero. Otra alternativa quizás más interesante, sería combinar las dos anteriores, de forma que se cobraría una prima de riesgo al tomador con la que poder financiar la transferencia del riesgo de impago, sin que suponga un coste adicional a la aseguradora.

En referencia a los indicadores macroeconómicos, cuya obtención no supone coste alguno al encontrarse en el INE, encontramos una notable mejora en la calidad de las predicciones de impago con su introducción; siendo el peso del sector secundario en el PIB provincial la variable que más información reporta, al dar una idea de su situación económica estructural, pues las barreras de entrada de estas actividades son elevadas. Además, la variación de la tasa de paro también se muestra como una variable a considerar en el análisis, ya que, junto con la variación del PIB, es la magnitud más empleada para evaluar la coyuntura económica de una región.

Atendiendo a los medios de pago se ha encontrado cómo las pólizas domiciliadas aumentan las probabilidades de impago, frente al efecto reductor que supone el pago anualizado respecto al que se encuentra fraccionado. Cabe destacar que, dada la tendencia que sigue el mercado orientada hacia la domiciliación bancaria, se deben buscar vías que compensen el aumento de casos de impago que se estiman van a ocurrir. Como se ha comprobado, el impacto positivo de anualizar el pago de la prima es mayor que el provocado por la domiciliación, por lo que podría ser una interesante vía para su compensación. A tal efecto, y en base a una cartera simulada con características parecidas a una real, se ha calculado el efecto positivo en términos de aumento de ingresos por disminución de probabilidades de impago, tras la modificación de las formas de pago en una serie de pólizas tomadas aleatoriamente como referencia.

Tras este análisis, se han presentado algunos aspectos que podrían tratarse en un futuro para perfeccionar los métodos propuestos, como son la introducción de técnicas de machine learning con el ejemplo del árbol de clasificación y el árbol de GLM, que sin ser muy refinados muestran leves mejoras frente a los iniciales; la elaboración de modelos GLM sintéticos, capaces de integrar como variable explicativa un modelo de riesgo crediticio puro asociado a sus principales ratios financieros; o el creciente desarrollo de la economía conductual, orientada a analizar las implicaciones que tiene el comportamiento de individuos y empresas en la toma de decisiones y sus consecuencias en términos de siniestralidad y propensión al impago, en el caso de estudio.

A modo de conclusión global, este estudio ha sido capaz de mostrar el potencial de la información que las entidades aseguradoras poseen y cómo puede ser tratada para la elaboración de modelos que permitan conocer antes del momento de emisión de una nueva póliza o del proceso de renovación de la cartera, las probabilidades de impago que llevan asociadas; dotándolas de una calificación que informará, en términos globales, cuántas pólizas se espera que no paguen la prima para cada rating. Con todo ello, la entidad dispondrá de la información necesaria para poder llevar a cabo estrategias orientadas a la reducción de su exposición mediante técnicas de cobertura y mitigación, de compensación basadas en el cobro de una prima de riesgo o de adaptación de las condiciones de contratación como podría ser el pago anualizado. Además, se ha comprobado como los factores macroeconómicos han de ser tenidos en cuenta para este tipo de valoraciones.

7. Glosario

- Lapses: denominación que recibe el conjunto de pólizas cuyos tomadores deciden unilateralmente, en la fecha de vencimiento, la no renovación del contrato de seguro.
- Producto Interior Bruto (PIB): valor monetario de todos los bienes y servicios de demanda final producidos en una región durante un período determinado de tiempo. Es considerada la principal magnitud macroeconómica.
- Modelo de Regresión: proceso estadístico cuyo objetivo es capturar y analizar las relaciones existentes entre una variable o conjunto de variables, denominadas explicativas, respecto a otra denominada dependiente u objetivo.
- Multicolinealidad: situación en la que dos o más variables explicativas de un modelo de regresión muestran una alta correlación. Esta se considerará perfecta cuando las variables se puedan presentar como combinaciones lineales de otras.
- Estimador: estadístico procedente de una muestra de datos empleado para la estimación de parámetros desconocidos de la población.
- Sesgo: asociado a un estimador, es la diferencia existente entre el valor real de un parámetro y el valor esperado calculado para el mismo.
- Consistencia: condición por la que el valor de un estimador converge al valor real del parámetro, cuando aumenta el tamaño de la muestra empleada para su cálculo.
- Matriz de Confusión[11]: herramienta empleada para la evaluación de un algoritmo de predicción para datos binarios o que responden ante una determinada característica, que contabiliza el número de predicciones que fueron correcta e incorrectamente clasificadas por el algoritmo frente a su valor real. A modo de ejemplo: en el caso binario dónde las observaciones sólo pueden tomar los valores $Impago = 1$ y $Pago = 0$, son cuatro los casos contemplados:
- Curva ROC: representa a través de una gráfica la relación existente entre la especificidad y la sensibilidad para un sistema de clasificación de variables binarias, en función del umbral de discriminación empleado. Su nombre procede del acrónimo “Receiver Operating Characteristic”.

[11] https://en.wikipedia.org/wiki/Confusion_matrix

- Seguro de Crédito[12]: garantiza a la persona o entidad asegurada el pago de aquellos créditos pendientes de cobro en el momento en que se produzca la insolvencia del deudor.
- Mercado Over The Counter (OTC): reciben también el nombre de mercados no organizados y se caracterizan, principalmente, por la negociación de instrumentos financieros y otros productos derivados en los que son las partes las que acuerdan las condiciones particulares de los contratos, no estando estandarizadas las operaciones que en estos se producen. Aunque inicialmente no existía un sistema de compensación en caso de default, la normativa que se está desarrollando en torno a ellos es cada vez mayor.
- Swap: denominada también permuta financiera, se basa en el intercambio de ciertas cantidades de dinero en la forma y plazos acordados entre las partes. Destacan las ligadas a canjes de sumas que se encuentran a tipos de interés variable por sumas a tipos de interés fijo.

[12] https://www.fundacionmapfre.org/fundacion/es_es/publicaciones/diccionario-mapfre-seguros/s/seguro-de-credito.jsp

8. Referencias Bibliográficas

Bases de Datos Utilizadas

Fuentes Internas de Entidad Aseguradora Nacional (2012 - 2017). Recuperado el 13 de Abril de 2018

Instituto Nacional de Estadística (2012 - 2017). Recuperado el 24 de Mayo de 2018, de http://www.ine.es

Documentos Consultados

Lyn C., Thomas. *A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers.* International Journal of Forecasting 16 (2000) 149-172

Crook, Jonathan N. et al. *Recent developments in consumer credit risk assessment*. European Journal of Operational Research 183 (2007) 1447-1465

Milborrow, Stephen (2018). *Plotting rpart tres with the rpart.plot package*

Marin, J.M. *Tema 3: Modelos lineales generalizados.* Universidad Carlos III Madrid

Rubio. José A. et al. (2017). *El sector del seguro, la transformación hacia el risk management integral y personalizado*. Minsait by Indra

Stock, James H. y Watson, Mark W. (2012). *Introducción a la Econometría.*

Siegel, Eric (2016). *Predictive Analytics*

Mueller, John P. y Massaron, Luca (2016). *Machine Learning for dummies*

Gareth, J. et al. (2013). *An Introduction to Statistical Learning with Applications in R*

Hastie, Trevor et al. (2008). *The Elements of Statistical Learning*

Correa, Juan C., y González, Nelfi (2002). *Gráficos Estadísticos con R*. Universidad Nacional-Sede Medellín

Powers, David M. W. (2007). *Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation*. Flinders University of South Australia

Miranda, Aloysio (2003). *El Método de Remuestreo y su Aplicación en la Investigación Biomédica*. Escuela Nacional de Salud Pública "Carlos J. Finlay"

Páginas Web Visitadas

Marsh LLC. *Global Insurance Market Index – First Quarter 2018.* Recuperado el 01 de Junio de 2018, de www.marsh.com/sg/insights/research/

DATA IKU. *Machine Learning Basics – An Illustrated Guide for Non-Technical Readers*. Recuperado el 09 de Febrero de 2018, de https://pages.dataiku.com/

DATA IKU. *Building Production-Ready Predictive Analytics*. Recuperado el 09 de Febrero de 2018, de https://pages.dataiku.com/

Fundación MAPFRE*[13]. *Diccionario MAPFRE de seguros – Una guía en el mundo del seguro. Recuperado de* www.fundacionmapfre.org/fundacion/es_es/publicaciones/diccionario-mapfre-seguros/

Colaboradores de Wikipedia*. *Wikipedia, La enciclopedia libre*. Recuperado de https://es.wikipedia.org/wiki

Real Academia Española*. *Diccionario de lengua española.* Recuperado de http://dle.rae.es/

Milborrow, Stephen (2011). *Plot A Model´s Response With Varying Predictor Values.* Recuperado el 28 de Mayo de 2018, de https://www.rdocumentation.org/packages/plotmo/versions/1.0-0/topics/plotmo

Mendoza, Juan B. (2018). *Árboles de decisión con R – Clasificación*. Recuperado el 24 de Mayo de 2018, de https://rpubs.com/jboscomendoza/arboles_decision_clasificacion

Expansión. Economía. Recuperado el 18 de Mayo de 2018, de http://www.expansion.com/economia/2017/10/09/59db4970e2704e82778b45ee.html

Samson, Alain. *An Introduction to Behavioral Economics*. Recuperado el 18 de Mayo de 2018, de https://www.behavioraleconomics.com/introduction-behavioral-economics/

Stack Exchange. *Interpreting Residual and Null Deviance in GLM R.* Recuperado el 26 de Mayo de 2018, de https://stats.stackexchange.com/questions/108995/interpreting-residual-and-null-deviance-in-glm-r

[13] El símbolo (*) indica que la web ha sido consultada en más de una ocasión.

9. Anexos

Anexo 1. Variables

Nombres asignados a cada variable dentro de los modelos estimados para su simplificación.

Variable	Denominación
Paga_Impaga	**Y1**
Obs	**X1**
Familia_Sector	**X4**
Mes_Origen	**X6**
Ln_Capital	**X9**
Coaseguro	**X11**
Reaseguro	**X12**
Ln_PrimaNeta	**X14**
Poliza_NP	**X16**
Antiguedad	**X20**
Canal	**X22**
Codigo_Territorial	**X23**
C1_Periodicidad_Pago	**X27**
C2_Periodicidad_Pago	**X28**
Domiciliacion	**X29**
Negocio_Especializado	**X31**

Variable	Denominación
CVM	**X32**
C_ClubEmpresa	**X33**
Agravación	**X34**
NumS_Cum	**X72**
Poliza_Anulada	**X78**
Vigor_Impago_Otro	**X80**
PAC	**X84**
DimEmp	**X87**
Paro	**X89**
Var_Paro	**X90**
S1	**X91**
S2	**X92**
S3	**X93**
M_ORIG12	**X94**
C_M_ORIG	**X95**
C_Trimestre	**X96**

Anexo 2. Modelos

A continuación se muestran las estimaciones realizadas para los coeficientes de los modelos logit y probit referenciados en el epígrafe de modelización.

2.1 –Modelo Inicial

Variables	Logit			Probit		
	Coeficiente	Std. Error		Coeficiente	Std. Error	
Intercepto	-2,184	0,123	***	-1,265	0,065	***
X4S2	0,013	0,045		0,017	0,024	
X4S3	0,068	0,025	**	0,042	0,014	**
X610	-0,074	0,051		-0,038	0,028	
X611	-0,046	0,051		-0,027	0,028	
X612	-0,109	0,050	*	-0,060	0,027	*
X62	-0,090	0,050	.	-0,049	0,027	.
X63	-0,004	0,049		-0,004	0,027	
X64	0,025	0,050		0,015	0,028	
X65	0,045	0,049		0,024	0,027	
X66	0,079	0,049		0,045	0,027	.
X67	0,020	0,049		0,010	0,027	
X68	-0,069	0,057		-0,038	0,031	
X69	-0,005	0,054		-0,003	0,030	
X9	0,112	0,006	***	0,063	0,003	***
X111	-0,360	0,393		-0,126	0,172	
X121	-0,776	0,060	***	-0,406	0,029	***
X14	-0,216	0,016	***	-0,123	0,008	***
X161	0,057	0,071		0,035	0,040	
X22BR	-0,094	0,087		-0,057	0,046	
X22C	-0,085	0,024	***	-0,050	0,013	***
X22D	0,460	0,452		0,276	0,267	
X22O	-0,361	0,055	***	-0,177	0,030	***
X2310	0,320	0,044	***	0,175	0,024	***
X232	0,058	0,039		0,034	0,021	
X233	0,228	0,037	***	0,127	0,020	***
X234	0,310	0,035	***	0,172	0,019	***
X235	0,538	0,037	***	0,300	0,020	***
X239	-1,221	0,129	***	-0,601	0,060	***
X27I	-2,393	0,583	***	-1,048	0,218	***
X27S	0,533	0,025	***	0,293	0,014	***
X27T	0,617	0,032	***	0,329	0,018	***
X291	0,243	0,035	***	0,125	0,019	***
X311	0,004	0,025		0,000	0,014	
X331	-0,353	0,025	***	-0,192	0,014	***
X72	-0,013	0,003	***	-0,003	0,001	*
Significatividad: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

2.2 – Repercusión de Actividad Empresarial

Estimaciones obtenidas si se mantiene u omite la actividad empresarial del asegurado.

2.2.1 - Incluyendo el Sector de Actividad (Modelo 4)

Variables	Logit			Probit		
	Coeficiente	Std. Error		Coeficiente	Std. Error	
Intercepto	-1,839	0,111	***	-1,104	0,059	***
X4S2	-0,005	0,044		0,007	0,024	
X4S3	0,020	0,025		0,020	0,014	
X9	0,097	0,006	***	0,055	0,003	***
X121	-0,608	0,061	***	-0,322	0,030	***
X14	-0,184	0,014	***	-0,101	0,007	***
X202	-0,394	0,025	***	-0,220	0,014	***
X203	-0,610	0,030	***	-0,332	0,017	***
X22BR	-0,262	0,086	**	-0,133	0,045	**
X22C	-0,100	0,024	***	-0,056	0,013	***
X22D	0,294	0,436		0,167	0,250	
X22O	-0,302	0,055	***	-0,139	0,030	***
X2310	0,297	0,044	***	0,162	0,024	***
X232	0,068	0,039	.	0,039	0,021	.
X233	0,190	0,038	***	0,108	0,020	***
X234	0,281	0,035	***	0,155	0,019	***
X235	0,487	0,037	***	0,270	0,021	***
X239	-1,245	0,128	***	-0,599	0,060	***
X27I	-2,228	0,454	***	-1,019	0,179	***
X27S	0,497	0,025	***	0,272	0,014	***
X27T	0,562	0,032	***	0,295	0,018	***
X291	0,238	0,035	***	0,124	0,019	***
X331	-0,362	0,025	***	-0,197	0,014	***
X962	0,110	0,029	***	0,060	0,016	***
X963	0,056	0,030	.	0,028	0,017	.
X964	-0,006	0,029		-0,006	0,016	

Significatividad: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

2.2.2 - Omitiendo el Sector de Actividad (Modelo 3)

Variables	Logit			Probit		
	Coeficiente	Std. Error		Coeficiente	Std. Error	
Intercepto	-1,818	0,109	***	-1,084	0,058	***
X9	0,097	0,006	***	0,055	0,003	***
X121	-0,611	0,061	***	-0,324	0,030	***
X14	-0,185	0,014	***	-0,102	0,007	***
X202	-0,396	0,025	***	-0,221	0,014	***
X203	-0,612	0,030	***	-0,333	0,017	***
X22BR	-0,259	0,086	**	-0,130	0,045	**
X22C	-0,100	0,024	***	-0,056	0,013	***
X22D	0,296	0,436		0,169	0,250	
X22O	-0,302	0,055	***	-0,138	0,030	***
X2310	0,298	0,044	***	0,163	0,024	***
X232	0,069	0,039	.	0,041	0,021	.
X233	0,191	0,037	***	0,109	0,020	***
X234	0,283	0,035	***	0,157	0,019	***
X235	0,489	0,037	***	0,273	0,020	***
X239	-1,252	0,128	***	-0,606	0,059	***
X27I	-2,227	0,454	***	-1,019	0,179	***
X27S	0,496	0,025	***	0,272	0,014	***
X27T	0,561	0,032	***	0,295	0,018	***
X291	0,238	0,035	***	0,123	0,019	***
X331	-0,362	0,025	***	-0,197	0,014	***
X962	0,110	0,029	***	0,060	0,016	***
X963	0,056	0,030	.	0,028	0,017	.
X964	-0,006	0,029		-0,006	0,016	

Significatividad: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

2.3 – Pago Único frente al Fraccionado

Extensión del Modelo 3 modificando la forma de análisis del fraccionamiento del pago.

Variables	Logit			Probit		
	Coeficiente	Std. Error		Coeficiente	Std. Error	
Intercepto	-1,443	0,114	***	-0,880	0,061	***
X4S2	-0,002	0,044		0,007	0,024	
X4S3	0,019	0,025		0,019	0,014	
X9	0,097	0,006	***	0,055	0,003	***
X121	-0,606	0,061	***	-0,321	0,030	***
X14	-0,174	0,014	***	-0,097	0,007	***
X202	-0,376	0,025	***	-0,210	0,014	***
X203	-0,593	0,030	***	-0,321	0,017	***
X22BR	-0,252	0,086	**	-0,128	0,045	**
X22C	-0,097	0,024	***	-0,055	0,013	***
X22D	0,303	0,435		0,171	0,250	
X22O	-0,287	0,055	***	-0,134	0,030	***
X2310	0,298	0,044	***	0,162	0,024	***
X232	0,065	0,039	.	0,038	0,021	.
X233	0,188	0,038	***	0,107	0,020	***
X234	0,281	0,035	***	0,155	0,019	***
X235	0,486	0,037	***	0,270	0,020	***
X239	-1,245	0,127	***	-0,607	0,059	***
X28U	-0,501	0,022	***	-0,271	0,012	***
X291	0,264	0,035	***	0,136	0,018	***
X331	-0,360	0,025	***	-0,196	0,014	***
X962	0,108	0,029	***	0,059	0,016	***
X963	0,052	0,030	.	0,026	0,017	
X964	-0,011	0,029		-0,009	0,016	
Significatividad: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

2.4 –Variables Macroeconómicas

Estimaciones obtenidas incluyendo los indicadores macroeconómicos en los distintos modelos previamente estimados.

2.4.1 – Ampliación del Modelo 1

Variables	Logit			Probit		
	Coeficiente	Std. Error		Coeficiente	Std. Error	
Intercepto	3,302	0,208	***	1,855	0,115	***
X4S2	0,006	0,045		0,014	0,024	
X4S3	0,085	0,026	**	0,053	0,014	***
X610	-0,074	0,052		-0,037	0,028	
X611	-0,040	0,051		-0,023	0,028	
X612	-0,095	0,050	.	-0,052	0,027	.
X62	-0,084	0,050	.	-0,046	0,027	.
X63	-0,001	0,050		-0,004	0,027	
X64	0,025	0,051		0,013	0,028	
X65	0,035	0,050		0,018	0,028	
X66	0,088	0,049	.	0,047	0,027	.
X67	0,026	0,049		0,013	0,027	
X68	-0,097	0,058	.	-0,054	0,032	.
X69	-0,007	0,055		-0,004	0,030	
X9	0,127	0,006	***	0,071	0,003	***
X111	-0,383	0,395		-0,124	0,170	
X121	-0,713	0,060	***	-0,369	0,029	***
X14	-0,230	0,016	***	-0,130	0,008	***
X161	0,077	0,072		0,051	0,040	
X22BR	-0,155	0,088	.	-0,085	0,047	.
X22C	-0,074	0,024	**	-0,044	0,013	***
X22D	0,340	0,457		0,210	0,269	
X22O	-0,483	0,057	***	-0,232	0,031	***
X2310	-0,033	0,046		-0,030	0,025	
X232	-0,070	0,040	.	-0,042	0,021	*
X233	-0,106	0,040	**	-0,063	0,022	**
X234	-0,146	0,041	***	-0,079	0,023	***
X235	-0,504	0,053	***	-0,285	0,030	***
X239	-1,482	0,133	***	-0,743	0,063	***
X27I	-2,470	0,584	***	-1,081	0,217	***
X27S	0,509	0,025	***	0,279	0,014	***
X27T	0,592	0,032	***	0,312	0,018	***
X291	0,263	0,036	***	0,133	0,019	***
X311	-0,011	0,025		-0,009	0,014	
X331	-0,341	0,025	***	-0,183	0,014	***
X72	-0,015	0,004	***	-0,003	0,002	*
X90	-0,171	0,581		-0,188	0,327	
X92	-8,377	0,302	***	-4,698	0,167	***
X93	-5,513	0,189	***	-3,150	0,107	***
Significatividad: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1						

2.4.2 – Ampliación del Modelo 2

Variables	Logit			Probit		
	Coeficiente	Std. Error		Coeficiente	Std. Error	
Intercepto	3,588	0,204	***	1,990	0,113	***
X4S2	-0,008	0,045		0,006	0,024	
X4S3	0,044	0,025	.	0,034	0,014	*
X9	0,112	0,006	***	0,063	0,003	***
X121	-0,545	0,062	***	-0,286	0,030	***
X14	-0,197	0,015	***	-0,109	0,008	***
X202	-0,385	0,026	***	-0,217	0,015	***
X203	-0,597	0,031	***	-0,325	0,017	***
X22BR	-0,298	0,087	***	-0,150	0,045	***
X22C	-0,087	0,024	***	-0,050	0,013	***
X22D	0,250	0,441		0,138	0,252	
X22O	-0,420	0,057	***	-0,194	0,031	***
X2310	-0,051	0,046		-0,042	0,025	
X232	-0,057	0,040		-0,036	0,022	.
X233	-0,135	0,040	***	-0,078	0,022	***
X234	-0,166	0,042	***	-0,092	0,023	***
X235	-0,539	0,053	***	-0,306	0,030	***
X239	-1,507	0,132	***	-0,740	0,062	***
X27I	-2,286	0,454	***	-1,032	0,178	***
X27S	0,475	0,025	***	0,259	0,014	***
X27T	0,539	0,032	***	0,280	0,018	***
X291	0,260	0,036	***	0,132	0,019	***
X331	-0,351	0,025	***	-0,189	0,014	***
X72	-0,004	0,003		-0,001	0,001	
X962	0,105	0,029	***	0,056	0,016	***
X963	0,045	0,031		0,023	0,017	
X964	-0,005	0,030		-0,003	0,016	
X90	-0,118	0,582		-0,140	0,328	
X92	-8,304	0,302	***	-4,659	0,167	***
X93	-5,487	0,190	***	-3,128	0,107	***
Significatividad: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

2.4.3 – Ampliación del Modelo 3

Variables	Logit			Probit		
	Coeficiente	Std. Error		Coeficiente	Std. Error	
Intercepto	3,628	0,202	***	2,000	0,112	***
X4S2	-0,007	0,045		0,006	0,024	
X4S3	0,041	0,025		0,033	0,014	*
X9	0,112	0,006	***	0,063	0,003	***
X121	-0,548	0,062	***	-0,287	0,030	***
X14	-0,203	0,014	***	-0,110	0,008	***
X202	-0,388	0,026	***	-0,218	0,015	***
X203	-0,603	0,030	***	-0,326	0,017	***
X22BR	-0,302	0,087	***	-0,151	0,045	***
X22C	-0,088	0,024	***	-0,050	0,013	***
X22D	0,252	0,441		0,139	0,252	
X22O	-0,421	0,057	***	-0,195	0,031	***
X2310	-0,050	0,046		-0,041	0,025	
X232	-0,057	0,040		-0,036	0,022	.
X233	-0,134	0,040	***	-0,078	0,022	***
X234	-0,166	0,042	***	-0,092	0,023	***
X235	-0,539	0,053	***	-0,306	0,030	***
X239	-1,511	0,132	***	-0,741	0,062	***
X27I	-2,288	0,454	***	-1,032	0,178	***
X27S	0,474	0,025	***	0,259	0,014	***
X27T	0,539	0,032	***	0,281	0,018	***
X291	0,261	0,036	***	0,133	0,019	***
X331	-0,351	0,025	***	-0,189	0,014	***
X962	0,105	0,029	***	0,056	0,016	***
X963	0,046	0,031		0,023	0,017	
X964	-0,004	0,030		-0,003	0,016	
X90	-0,119	0,582		-0,141	0,328	
X92	-8,299	0,302	***	-4,657	0,167	***
X93	-5,485	0,190	***	-3,127	0,107	***
Significatividad: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

2.4.4 – Ampliación del Modelo 4

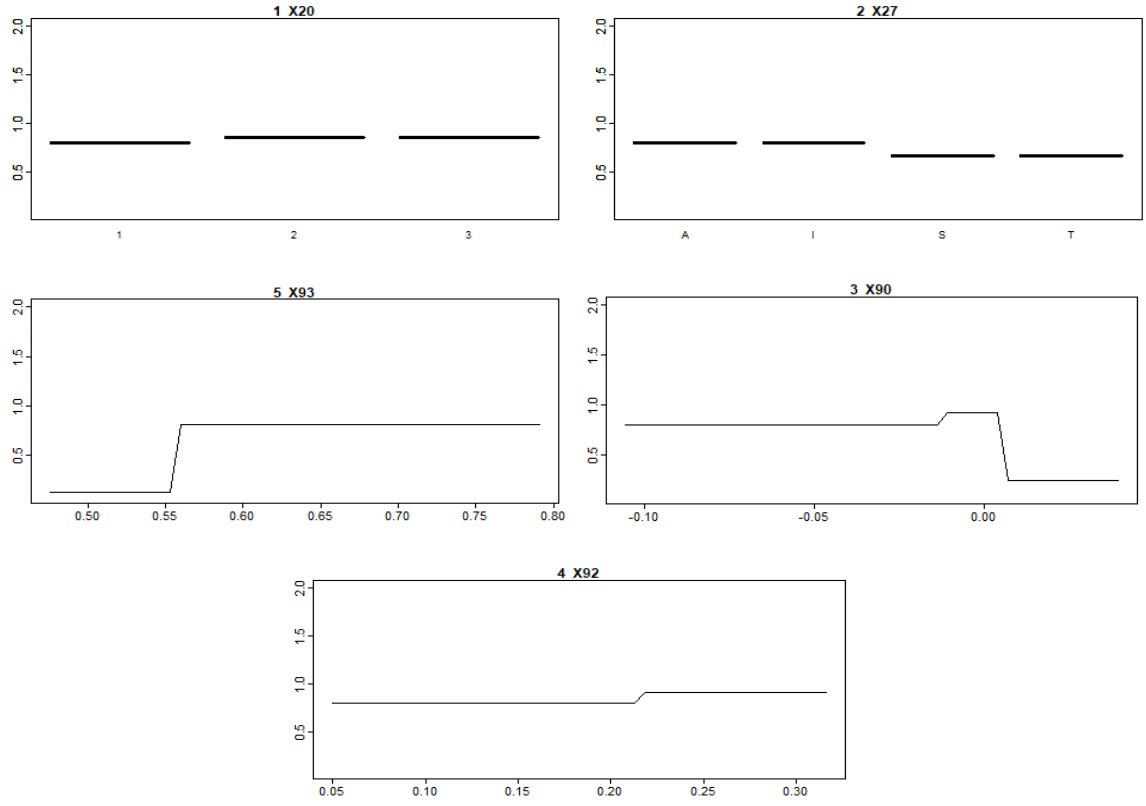
Variables	Logit			Probit		
	Coeficiente	Std. Error		Coeficiente	Std. Error	
Intercepto	3,657	0,201	***	2,021	0,112	***
X9	0,112	0,006	***	0,063	0,003	***
X121	-0,555	0,061	***	-0,292	0,030	***
X14	-0,206	0,014	***	-0,112	0,008	***
X202	-0,390	0,026	***	-0,220	0,015	***
X203	-0,606	0,030	***	-0,328	0,017	***
X22BR	-0,295	0,087	***	-0,147	0,045	**
X22C	-0,089	0,024	***	-0,051	0,013	***
X22D	0,254	0,441		0,141	0,252	
X22O	-0,419	0,057	***	-0,193	0,030	***
X2310	-0,049	0,046		-0,041	0,025	
X232	-0,056	0,040		-0,035	0,022	
X233	-0,134	0,040	***	-0,077	0,022	***
X234	-0,166	0,042	***	-0,092	0,023	***
X235	-0,536	0,053	***	-0,304	0,030	***
X239	-1,529	0,131	***	-0,753	0,061	***
X27I	-2,286	0,454	***	-1,032	0,178	***
X27S	0,474	0,025	***	0,259	0,014	***
X27T	0,539	0,032	***	0,280	0,018	***
X291	0,260	0,036	***	0,132	0,019	***
X331	-0,351	0,025	***	-0,189	0,014	***
X962	0,105	0,029	***	0,056	0,016	***
X963	0,046	0,031		0,023	0,017	
X964	-0,005	0,030		-0,003	0,016	
X90	-0,116	0,582		-0,140	0,328	
X92	-8,317	0,302	***	-4,670	0,167	***
X93	-5,456	0,189	***	-3,104	0,107	***
Significatividad: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

2.4.5 – Ampliación del Modelo 9

Variables	Logit			Probit		
	Coeficiente	Std. Error		Coeficiente	Std. Error	
Intercepto	3,947	0,205	***	2,196	0,114	***
X4S2	-0,006	0,045		0,006	0,024	
X4S3	0,043	0,025	.	0,033	0,014	*
X9	0,113	0,006	***	0,063	0,003	***
X121	-0,544	0,061	***	-0,285	0,030	***
X14	-0,187	0,015	***	-0,105	0,008	***
X202	-0,366	0,026	***	-0,207	0,015	***
X203	-0,579	0,031	***	-0,314	0,017	***
X22BR	-0,287	0,087	***	-0,145	0,045	**
X22C	-0,085	0,024	***	-0,049	0,013	***
X22D	0,258	0,441		0,142	0,252	
X22O	-0,403	0,056	***	-0,189	0,030	***
X2310	-0,049	0,046		-0,041	0,025	
X232	-0,061	0,040		-0,037	0,022	.
X233	-0,137	0,040	***	-0,079	0,022	***
X234	-0,166	0,041	***	-0,093	0,023	***
X235	-0,540	0,053	***	-0,306	0,030	***
X239	-1,506	0,130	***	-0,749	0,061	***
X28U	-0,478	0,023	***	-0,257	0,013	***
X291	0,287	0,035	***	0,145	0,019	***
X331	-0,349	0,025	***	-0,188	0,014	***
X72	-0,004	0,003		-0,001	0,001	
X962	0,103	0,029	***	0,055	0,016	***
X963	0,041	0,031		0,021	0,017	
X964	-0,010	0,030		-0,005	0,016	
X90	-0,168	0,582		-0,169	0,327	
X92	-8,290	0,302	***	-4,656	0,167	***
X93	-5,480	0,189	***	-3,126	0,107	***
Significatividad: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

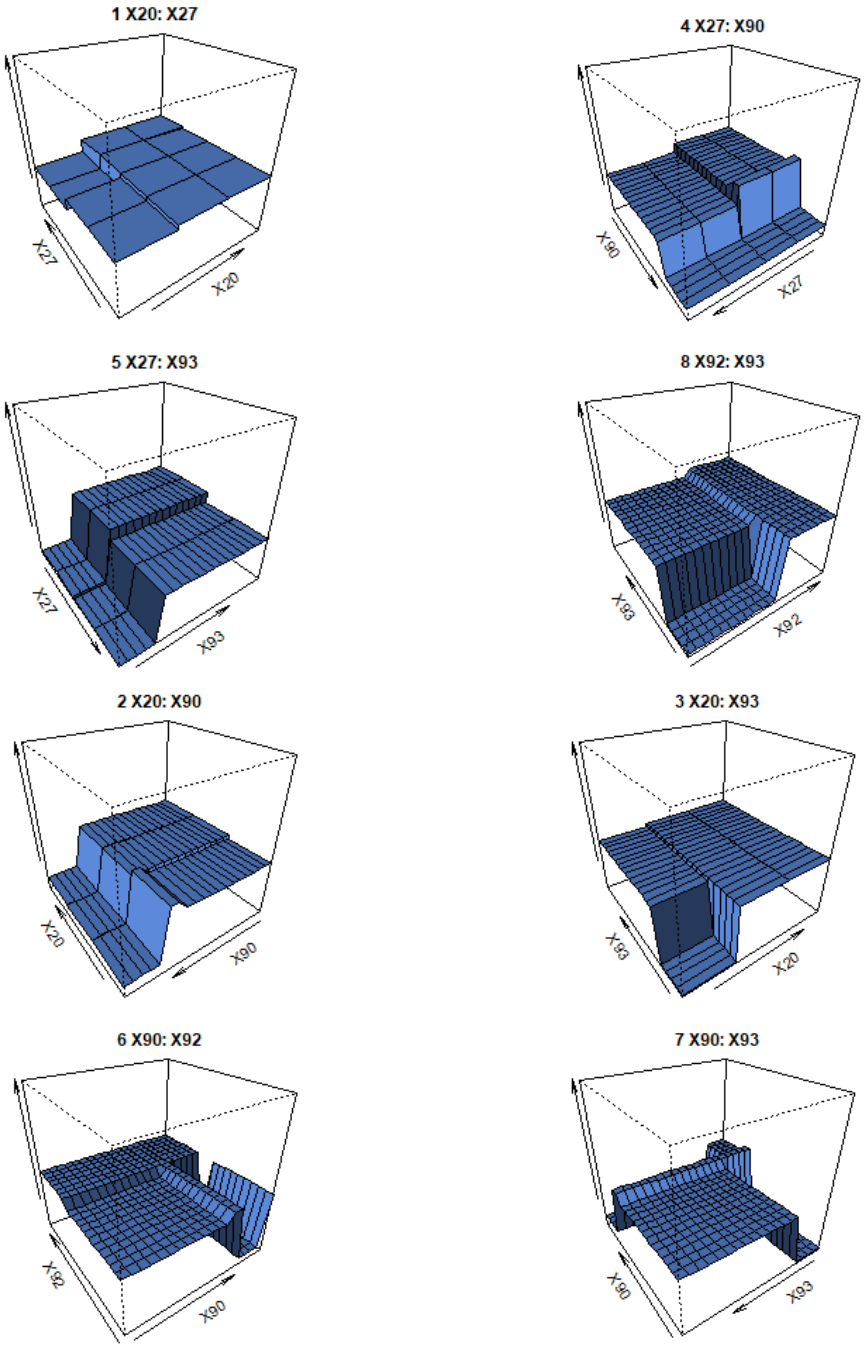
Anexo 3. Gráficos del Árbol de Clasificación

3.1 –Análisis Univariable



3.2 –Análisis Multivariable

Los gráficos mostrados recogen la relación existente entre la interacción de las variables macroeconómicas, la forma de pago y la antigüedad de la póliza en cartera respecto a la probabilidad de pago de la prima.



Siendo las variables:

- **X20:** Antigüedad de la Póliza
- **X27:** Modalidad de Pago
- **X90:** Tasa de Variación del Paro
- **X92 y X93:** Peso de los sectores Secundario y Terciario en el PIB Provincial

Anexo 4. Aplicación de GLMs sobre el Árbol de Clasificación

4.1 – Modelos Estimados

A continuación se recogen los coeficientes para cada una de las ramas del árbol de clasificación.

4.1.1-Rama 1

Variables	Logit			Probit		
	Coeficiente	Std. Error		Coeficiente	Std. Error	
Intercepto	9,749	0,629	***	4,938	0,331	***
X4S2	-0,027	0,097		-0,010	0,050	
X4S3	0,117	0,050	*	0,068	0,026	**
X9	0,088	0,014	***	0,046	0,007	***
X121	-0,216	0,118	.	-0,099	0,056	.
X14	-0,160	0,030	***	-0,078	0,015	***
X202	-0,395	0,056	***	-0,212	0,030	***
X203	-0,575	0,064	***	-0,301	0,034	***
X22BR	-0,022	0,187		0,033	0,092	
X22C	-0,042	0,049		-0,021	0,026	
X22D	0,417	1,145		0,231	0,626	
X22O	0,059	0,141		0,030	0,074	
X2310	-0,368	0,090	***	-0,190	0,050	***
X232	-0,681	0,076	***	-0,359	0,040	***
X233	-0,567	0,082	***	-0,305	0,045	***
X234	-1,055	0,086	***	-0,504	0,045	***
X235	-0,666	0,396	.	-0,288	0,201	
X239	-2,160	0,276	***	-1,012	0,127	***
X27I	-1,275	0,735	.	-0,485	0,306	
X27S	0,470	0,055	***	0,244	0,029	***
X27T	0,610	0,069	***	0,308	0,037	***
X291	0,384	0,071	***	0,185	0,036	***
X331	-0,220	0,051	***	-0,117	0,027	***
X962	0,091	0,060		0,047	0,032	
X963	0,008	0,065		0,000	0,034	
X964	-0,004	0,061		0,003	0,032	
X90	-17,295	1,432	***	-9,177	0,765	***
X92	-13,787	1,050	***	-7,170	0,553	***
X93	-13,520	0,715	***	-7,068	0,372	***
Significatividad: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

4.1.2-Rama 2

Variables	Logit			Probit		
	Coeficiente	Std. Error		Coeficiente	Std. Error	
Intercepto	1,442	0,279	***	0,732	0,154	***
X4S2	-0,070	0,061		-0,025	0,033	
X4S3	0,013	0,035		0,020	0,019	
X9	0,138	0,009	***	0,075	0,004	***
X121	-0,595	0,075	***	-0,312	0,037	***
X14	-0,238	0,019	***	-0,127	0,010	***
X203	-0,232	0,032	***	-0,122	0,017	***
X22BR	-0,157	0,112		-0,088	0,060	
X22C	-0,046	0,034		-0,028	0,019	
X22D	0,453	0,538		0,272	0,315	
X22O	-0,241	0,070	***	-0,106	0,038	**
X2310	0,041	0,071		0,017	0,039	
X232	0,631	0,066	***	0,347	0,036	***
X233	0,029	0,061		0,018	0,033	
X234	0,022	0,064		0,014	0,035	
X235	-0,159	0,076	*	-0,091	0,041	*
X239	-1,331	0,162	***	-0,651	0,077	***
X27I	-8,775	74,774		-3,034	21,699	
X27S	0,394	0,036	***	0,214	0,020	***
X27T	0,343	0,045	***	0,171	0,025	***
X291	0,298	0,049	***	0,154	0,026	***
X331	-0,385	0,036	***	-0,205	0,019	***
X962	0,020	0,040		0,012	0,022	
X963	-0,052	0,043		-0,029	0,023	
X964	-0,081	0,041	*	-0,045	0,023	*
X90	1,444	0,869	.	0,766	0,481	
X92	-5,115	0,540	***	-2,870	0,298	***
X93	-3,915	0,256	***	-2,183	0,143	***
Significatividad: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

4.1.3-Rama 3

Variables	Logit			Probit		
	Coeficiente	Std. Error		Coeficiente	Std. Error	
Intercepto	2,701	0,445	***	1,627	0,259	***
X4S2	0,059	0,104		0,035	0,061	
X4S3	-0,020	0,061		-0,016	0,036	
X9	0,046	0,014	**	0,026	0,008	**
X121	-0,502	0,416		-0,332	0,210	
X14	-0,208	0,034	***	-0,121	0,019	***
X22BR	-1,020	0,243	***	-0,578	0,128	***
X22C	-0,248	0,052	***	-0,147	0,030	***
X22D	-0,475	1,110		-0,341	0,623	
X22O	-1,193	0,143	***	-0,642	0,077	***
X2310	0,297	0,122	*	0,160	0,070	*
X232	0,592	0,118	***	0,332	0,068	***
X233	0,101	0,105		0,047	0,059	
X234	0,390	0,107	***	0,216	0,061	***
X235	0,105	0,120		0,047	0,069	
X239	-1,322	0,590	*	-0,570	0,269	*
X27I	-2,545	0,586	***	-1,166	0,227	***
X27S	0,666	0,054	***	0,392	0,032	***
X27T	0,848	0,071	***	0,491	0,042	***
X291	0,174	0,085	*	0,077	0,047	
X331	-0,415	0,056	***	-0,235	0,032	***
X962	0,258	0,062	***	0,148	0,036	***
X963	0,291	0,066	***	0,167	0,039	***
X964	0,197	0,064	**	0,119	0,037	**
X90	-1,309	1,349		-0,722	0,797	
X92	-4,901	0,838	***	-2,989	0,492	***
X93	-4,265	0,393	***	-2,508	0,233	***
Significatividad: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

4.1.14-Rama 4

Variables	Logit			Probit		
	Coeficiente	Std. Error		Coeficiente	Std. Error	
Intercepto	-14,250	4,260	***	-8,444	2,500	***
X4S2	-0,001	0,705		0,004	0,412	
X4S3	-0,200	0,349		-0,103	0,209	
X9	0,368	0,152	*	0,207	0,080	**
X14	-0,655	0,277	*	-0,376	0,160	*
X202	-0,181	0,467		-0,114	0,280	
X203	-0,195	0,510		-0,118	0,306	
X22BR	-0,713	1,512		-0,471	0,893	
X22C	-0,478	0,375		-0,294	0,224	
X22O	-18,310	957		-6,299	229,487	
X2310	0,431	1,854		0,294	1,120	
X232	-0,174	0,794		-0,150	0,469	
X233	1,308	1,126		0,802	0,682	
X234	0,179	0,617		0,119	0,371	
X235	-15,160	2344		-4,552	564,607	
X239	-15,290	3956		-4,230	973,499	
X27S	0,407	0,404		0,237	0,243	
X27T	1,290	0,585	*	0,774	0,343	*
X291	0,679	0,567		0,384	0,336	
X331	-0,309	0,435		-0,159	0,256	
X962	0,361	0,410		0,243	0,246	
X963	0,087	0,453		0,054	0,272	
X964	-0,981	0,514	.	-0,572	0,305	.
X90	-38,530	17,790	*	-24,193	10,607	*
X92	35,070	16,310	*	20,700	9,738	*
X93	12,750	3,861	***	7,712	2,296	***
Significatividad: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

4.1.5-Rama 5

Variables	Logit			Probit		
	Coeficiente	Std. Error		Coeficiente	Std. Error	
Intercepto	2,278	1,643		1,329	0,986	
X4S2	0,870	0,336	**	0,528	0,201	**
X4S3	0,035	0,184		0,023	0,112	
X9	0,236	0,065	***	0,140	0,034	***
X121	-1,103	0,476	*	-0,651	0,282	*
X14	-0,231	0,110	*	-0,135	0,064	*
X202	-0,147	0,196		-0,087	0,120	
X203	-0,316	0,235		-0,190	0,143	
X22BR	-0,161	0,508		-0,101	0,311	
X22C	-0,250	0,170		-0,159	0,104	
X22D	-16,537	2399,545		-5,189	376,754	
X22O	-1,921	0,422	***	-1,098	0,238	***
X2310	0,633	0,738		0,385	0,449	
X232	0,448	0,696		0,276	0,423	
X233	1,406	0,691	*	0,865	0,412	*
X234	0,466	0,576		0,288	0,348	
X235	-0,179	0,590		-0,088	0,357	
X239	-13,973	678,241		-3,777	105,987	
X27I	-15,779	622,224		-4,820	97,303	
X27S	0,488	0,189	**	0,302	0,115	**
X27T	0,898	0,214	***	0,543	0,130	***
X291	0,336	0,272		0,181	0,162	
X331	-0,489	0,177	**	-0,312	0,107	**
X962	0,286	0,201		0,188	0,122	
X963	0,206	0,209		0,124	0,127	
X964	0,152	0,218		0,096	0,132	
X90	-35,664	9,769	***	-21,397	5,875	***
X92	-11,110	3,560	**	-6,441	2,141	**
X93	-3,885	1,389	**	-2,317	0,835	**
Significatividad: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

4.2 – Fallos por Rama

Los modelos recogidos en el punto anterior dan lugar a las siguientes tasas de fallo en la previsión del impago.

	Tasa de Fallos Por Rama									
	Rama 1		Rama 2		Rama 3		Rama 4		Rama 5	
	Logit	Probit	Logit	Probit	Logit	Probit	Logit	Probit	Logit	Probit
40%	96,1%	97,8%	99,7%	100,0%	78,4%	79,3%	18,2%	18,2%	28,1%	28,1%
30%	86,2%	89,2%	95,7%	96,7%	51,6%	51,2%	11,4%	9,1%	12,4%	12,4%
20%	67,9%	67,3%	65,6%	65,0%	22,2%	21,7%	2,3%	2,3%	2,0%	2,0%
15%	54,6%	53,8%	38,1%	37,1%	10,5%	10,2%	2,3%	2,3%	2,0%	2,0%
10%	34,6%	33,2%	12,6%	12,0%	2,0%	2,2%	2,3%	2,3%	1,3%	1,3%
5%	9,1%	9,0%	2,0%	2,1%	0,6%	0,7%	0,0%	0,0%	0,7%	0,7%

Anexo 5. Formulación de la Matriz de Confusión[14]

A continuación se recoge la definición y formulación de los distintos ratios calculados a partir de las matrices de confusión de los modelos.

- Verdadero Positivo (VP): observación sobre la que se predijo el impago, siendo este su valor real.
- Verdadero Negativo (VN): observación sobre la que se predijo el pago, siendo este su valor real.
- Falso Positivo (FP): observación sobre la que se predijo el impago y finalmente pagó.
- Falso Negativo (FN): observación sobre la que se predijo el pago y finalmente incurrió en impago.
- Sensibilidad (VPR): capacidad para predecir los casos de impago, del total de impagos producidos.

$$VPR = \frac{VP}{VP + FN}$$

- Ratio de Falsos Positivos (FPR): probabilidad de que se estime como impago una observación que realmente pagó.

$$VPR = \frac{FP}{FP + VN}$$

[14] http://www2.cs.uregina.ca/~dbd/cs831/notes/confusion_matrix/confusion_matrix.html

- Exactitud (ACC): proporción del total de predicciones que fueron correctamente estimadas.

$$ACC = \frac{VP + VN}{VP + VN + FP + FN}$$

- Especificidad (SPC): capacidad para predecir los casos de pago, del total de pagos producidos.

$$SPC = \frac{TP}{TP + FN}$$

- Valores Predictivos Positivos (PPV): probabilidad de que impague si realmente no pago.

$$PPV = \frac{VP}{FP + VP}$$

- Valores Predictivos Negativos (NPV): probabilidad de que pague si realmente se produjo el pago.

$$NPV = \frac{VN}{VN + FN}$$

- Ratio de Falsos Descubrimientos (FRD): probabilidad de que estime incorrectamente que va a incurrir en impago, respecto al total de impagos.

$$FDR = \frac{FP}{FP + VP}$$